Bayesian Hierarchical Models to Address Problems in Neuroscience and Economics

by

Azeem Zaman

Program in Statistical and Economic Modeling
Duke University

Date: _____

Approved:

_____
Surya Tokdar, Supervisor

_____
Sayan Mukherjee

_____
Michelle Connolly

Thesis submitted in partial fulfillment of the requirements for the degree of
Master of Science in the Program in Statistical and Economic Modeling
in the Graduate School of Duke University
2017

Abstract

Bayesian Hierarchical Models to Address Problems in Neuroscience and Economics

by

Azeem Zaman

Program in Statistical and Economic Modeling
Duke University

Date: _____

Approved:

_____
Surya Tokdar, Supervisor

_____
Sayan Mukherjee

_____
Michelle Connolly

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Master of Science in the Program in Statistical and Economic Modeling
in the Graduate School of Duke University
2017

Abstract

In the first chapter, motivated by a model used to analyze spike train data, we present a method for learning multiple probability vectors by using information from large samples to improve estimates for smaller samples. The method makes use of Pólya-gamma data augmentation to construct a conjugate model whose posterior can estimate the weights of a mixture distribution. This novel method successfully uses borrows information from large samples to increase the precision and accuracy of estimates for smaller samples.

In the second chapter, data from the Federal Communications Commission spectrum auction number 73 is analyzed. By analyzing the structure of the auctions bounds are placed on the valuations that govern the bidders' decisions in the auction. With these bounds, common models are estimated by imputing valuations and the results are compared with the estimates from standard methods used in the literature. The comparison shows some important differences between the approaches. A second model that accounts for the geographic relationship between the licenses sold finds strong evidence of a correlation between the value of adjacent licenses, as expected by economic theory.

To Hanyu

Contents

List of Tables

List of Figures

List of Abbreviations and Symbols

Symbols

Here are the symbols used to denote common probabilistic operators and notions.

P	A probability measure
E	Expectation of a random variable
No	The normal distribution. All distributions are abbreviated using serif fonts. All referenced distributions are common, with the exception of the Pólya-gamma distribution, which is denoted PG.

Acknowledgements

I would like to thank my advisors for their excellent guidance. Without their help this thesis would be *far* worse.

1

Introduction

In this dissertation we present new Bayesian methods to address applied problems. The first of these problems is from Neuroscience.

This experiment is designed to analyze the firing rate of a neuron while it is exposed to two simultaneous auditory stimuli. In particular, a monkey listens to two different sounds from different locations and attempts to identify the origins. The impulses sent by a specific neuron are measured using an electrode inserted into the monkey's brain. The monkey is also exposed to both sounds (from the same locations that they are later emitted) individually. So we know how the neuron fires under sound A, sound B, and sound AB. Our goal is to model the firing rate AB as resulting from A and B in some way. Numerous methods for this process were hypothesised, and many different results were observed over the course of many replications of the experiment. For example, the neuron could ignore one sound, so that A is essentially the same as AB; the neuron could average A and B (perhaps in a time dependent manner) to generate AB; or the neuron could switch between A and B so that AB is sometimes similar to A and other times similar to B. As these are all time dependent measurements, the problem demands a model with great

flexibility to account for a wide variety of possible AB patterns.

Next we outline the essential features of the model in order to motivate the our new developments. We follow the presentation given by Glynn (2016). In the model we bin the data into T bins indexed by $t \in \{1, \ldots, T\}$. There are N replicate AB trails that are indexed by $i \in \{1, \ldots, N\}$. We observe $X_{i,t}$, which is the spike count for bin t in trail i. We assume that $X_{i,t}$ can be decomposed into response due to stimulus A and B. So $X_{i,t} = A_{i,t} + B_{i,t}$, with $A_{i,t}$ representing the number of spikes in response to A and $B_{i,t}$ representing the response to B. We observe $X_{i,t}$, but neither $A_{i,t}$ nor $B_{i,t}$. Next we introduce latent variable $A^*_{i,t}$ and $B^*_{i,t}$ for which

$$A_{i,t} \mid A^*_{i,t}, \alpha_{i,t} \sim \mathsf{Bi}(A^*_{i,t}, \alpha_{i,t})$$

for a trail and time specific mixing rate $\alpha_{i,t}$. The distribution for α is explained below. The latent variable $A^*_{i,t}$ is a Poisson random variable with *time specific mean* λ^A_t:

$$\begin{aligned} A^*_{i,t} \mid \lambda^A_t &\sim \mathsf{Po}(\lambda^A_t) \\ \lambda^A_t &\sim \mathsf{Ga}(r^A_t, s^A_t). \end{aligned}$$

The prior parameters for λ^A_t are determined from the observed counts for the single A trails. Analogously for $B_{i,t}$ we have

$$\begin{aligned} B_{i,t} \mid B^*_{i,t}, \alpha_{i,t} &\sim \mathsf{Bi}(B^*_{i,t}, 1 - \alpha_{i,t}) \\ B^*_{i,t} \mid \lambda^B_t &\sim \mathsf{Po}(\lambda^B_t) \\ \lambda^B_t &\sim \mathsf{Ga}(r^B_t, s^B_t).. \end{aligned}$$

Combining these results the observed distribution is

$$X_{i,t} \mid \lambda^A_t, \lambda^B_t, \alpha_{i,t} \sim \mathsf{Po}(\lambda^A_t \alpha_{i,t} + \lambda^B_t (1 - \alpha_{i,t})),$$

which means that the mean $X_{i,t}$ is a time-dependent convex combination of λ^A_t and λ^B_t. When $\alpha_{i,t} = 1$, the expectation is the same as what we would expect to observe if

only the A sound was present. We want $\alpha_{i,t}$ to be able to move very flexibly between 0 and 1 over time to capture a wide variety of potential neuron behavior. To achieve this, we model $\alpha_{i,t}$ as a logistic-Normal random process:

$$\alpha_{i,t} = \frac{e^{\eta_{i,t}}}{1 + e^{\eta_{i,t}}}$$

with $\eta_{i,t}$ representing the log odds of a spike occurring at time t for trial i. The trail log odds vector $\eta_{i,1:T}$ is distributed as a multivariate normal:

$$\begin{aligned} \eta_{i,1:T} &\sim \mathsf{No}(\bar{\eta}_i, C_i) \\ C_i(s,t) &= \sigma^2 K_\ell \\ K_{\ell_i} &= \exp\left(-\frac{|s-t|^2}{2\ell_i^2}\right). \end{aligned}$$

Therefore we have a Gaussian process with a squared exponential kernel. The length scale for trail i is modeled with a categorical distribution:

$$\begin{aligned} \ell_i \mid \rho &\sim \mathsf{Cat}((l_1, \ldots, l_L), \rho) \\ \rho &\sim \mathsf{Dir}(a). \end{aligned}$$

The set of possible length scales $(l_1, \ldots, l_L)$ is set in advance. A small value of l_i indicates a function that is able to shift rapidly between λ_t^A and λ_t^B, whereas a large value for l_i indicates that $\alpha_{i,t}$ is relatively flat. Note that probability vector ρ is shared across trails, so if the $\alpha_{i,t}$ curves should have a similar amount of "waviness" for all trails.

1.0.1 A Deficiency with the Model

After testing the model with both experimental and real data (see Caruso et al. (2017)), we see that the model generally performs well. From synthetic data trails, we know that the $\alpha_{i,t}$ curves are sufficiently flexible to model a wide variety of behavior. One problem with the model is that there is not enough information to

learn the distribution of ρ. We repeatedly observe that the posterior distribution of ρ is essentially the same as the prior.

The goal of this chapter is to develop a method to allow us to combine different experiments in a principled manner to learn the distribution of ρ. After stripping away the superfluous details of the model, we want to learn two probability vectors jointly. To approach this problem, we will attempt to learn the weights of two mixture distributions.

1.1 Frequentist Estimation of Mixture Weights

To extend the current model, we are interested learning the distribution of mixture weights in circumstances where we have two similar distributions. Specifically, suppose that $X_1 \sim f_1$ and $X_2 \sim f_2$, where

$$f_i(x) = \sum_{j=1}^{k} \phi_{j,i} g(x \mid \theta_j)$$

where θ_j is some known, fixed parameter and $\phi_{\mathbf{i}} = (\phi_{1,i}, \ldots, \phi_{k,i})$ is an unknown probability vector. We are interested in cases where ϕ_1 and ϕ_2 are similar, but not identical. To put it another way, we are interesting in "sharing" the information between X_1 and X_2. We want to do this without pooling the data, where we pretend $f_1 = f_2$ and $\phi_1 = \phi_2$. This sharing is designed to deal with a situation where X_1 or X_2 is a small sample, which will reduce the accuracy of our estimates. By sharing between the samples, we may be able to increase the accuracy of both estimates.

1.1.1 Maximum Likelihood Estimation

The first approach to estimate ϕ_i using a maximum likelihood estimate, which has a number of advantages. First of all, maximum likelihood is a simple technique that can be used for any f where we know how to evaluate $g(x \mid \theta_j)$. Secondly, there are many established results of the MLE. Specifically, under fairly mild regularity

conditions on g, we know that the MLE is asymptotically normal, which will allow us to calculate approximate standard errors for our estimates. The downside of the MLE is that we are not able to share between the estimates unless we are willing to pool them. In other words, ϕ_1 and ϕ_2 are estimated independently. If we know that the values of the two vectors should be similar, we are unable to use this to improve our estimates.

To perform maximum likelihood estimation, we will introduce the stick breaking map.

1.1.2 Stick breaking

Stick breaking is a technique often used in Bayesian nonparametrics to generate random discrete distributions. We will use the stick breaking map to map arbitrary vectors into probability vectors. If we denote the standard logistic function as

$$\Lambda(x) = \frac{1}{1 + e^{-x}} \tag{1.1}$$

then stick breaking is

$$\pi_{SB}(\mathbf{x}) = (\pi_1, \pi_2, \ldots, \pi_n), \tag{1.2}$$

where

$$\begin{aligned}
\pi_1 &= \Lambda(x_1) \\
\pi_j &= \Lambda(x_j) \prod_{i<j} (1 - \Lambda(x_i)) \\
\pi_n &= 1 - \sum_{i=1}^{n-1} \pi_i.
\end{aligned}$$

This maps any n dimensional vector $\mathbf{x}$ to an n dimensional probability vector. Note that the final component of $\mathbf{x}$, x_n, does not affect the result. For clarity, let $\bar{\pi}_{SB}$ denote the stick breaking map sending an n-dimensional $\mathbf{x}$ to a $n + 1$ dimensional

probability vector. This will be useful as it removes a totally free parameter from some optimization problems. We use the stick breaking construction because of its connection to the Pólya-gamma distribution, which will be introduced later.

1.1.3 Stick breaking and Maximum Likelihood

Using the stick breaking construction we can formulate the MLE. The MLE can be estimated as

$$\begin{aligned}\hat{\phi} &= \underset{\phi \in \Delta}{\operatorname{argmax}} \prod_{i=1}^{n} f(x_i) \\ &= \underset{\phi \in \Delta}{\operatorname{argmax}} \prod_{i=1}^{n} \sum_{m=1}^{k} \phi_m g(x_i \mid \theta_m),\end{aligned}$$

which is a constrained maximization problem. The constrain is $\phi \in \Delta$, where Δ is the probability simplex. The components of ϕ must sum to one and be non-negative. We can rewrite this as an unconstrained optimization problem using stick breaking:

$$\hat{\phi} = \bar{\pi}_{SB}\left(\underset{\mathbf{x} \in \mathbb{R}^{k-1}}{\operatorname{argmax}} \prod_{i=1}^{n} \sum_{m=1}^{k} \pi_m g(x_i \mid \theta_m)\right) \tag{1.3}$$

where π_m is the m component of $\bar{\pi}_{SB}(\mathbf{x})$. We are solving an unconstrained optimization problem, picking the vector in $\mathbb{R}^{k-1}$ such that stick breaking the vector gives the best possible probability vector. Using the invariance property of the MLE, we can then get an estimate for ϕ. Note that we are finding a vector $\mathbf{x} \in \mathbb{R}^{k-1}$. Stick breaking this $k-1$ dimensional vector gives us a k-dimensional probability vector. Finding a vector in $\mathbb{R}^k$ would result in the last component being completely free and preventing the existence of a unique solution.

One disadvantage of this approach is that the asymptotic variance become complicated. When we solve the constrained optimization problem, the Hessian gives the observed Fisher information. Taking the inverse of this matrix gives the asymptotic variance estimate. Solving the unconstrained problem and taking the inverse of the

Hessian gives us the variance of $\mathbf{x}$, from which we must derive the variance of the components of ϕ.

1.2 Bayesian Estimation of Mixture Weights

Suppose we have two probability vectors $\psi_i = (\psi^1, \ldots, \psi^k)$ for $i = 1, 2$. We observe samples

$$X_i \sim \sum_{j=1}^{k} \theta_i^j f(x \mid \theta_j)$$

where θ_j is a known vector of parameters for $j = 1, \ldots, k$. We wish to learn ψ_i in a Bayesian model. Instead of trying to learn ψ_i, suppose we try to learn $\mathbf{Y}_i = \pi_{SB}^{-1}(\psi_i)$, the vector Y_i such that when we apply the stick breaking map we get ψ_i. If the desired goal is to learn two similar vectors ψ_1 and ψ_2, we can translate this task to learning Y_1 and Y_2.

So if we want Y_1 and Y_2 to be "similar" in some way, we can accomplish this by inducing correlation between these vectors. The natural distribution to induce correlation between vectors is a multivariate normal. This choice is appropriate because because the distribution has full support, which is necessary to achieve all possible probability vectors. If the support of the distribution of Y was bounded, then the maximum and minimum values of π_1 are strictly less than 1 and strictly greater than 0. As an example, suppose that $k = 3$. We would have

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim \mathsf{No}\left(\mathbf{0}, \Sigma\right),$$

where

$$\Sigma = \begin{pmatrix} 1 & 0 & 0 & \rho_1 & 0 & 0 \\ 0 & 1 & 0 & 0 & \rho_2 & 0 \\ 0 & 0 & 1 & 0 & 0 & \rho_3 \\ \rho_1 & 0 & 0 & 1 & 0 & 0 \\ 0 & \rho_2 & 0 & 0 & 1 & 0 \\ 0 & 0 & \rho_3 & 0 & 0 & 1 \end{pmatrix}. \tag{1.4}$$

This design forces the first components of Y_1 and Y_2 to be correlated, which in turn ties together ψ_1^1 and ψ_2^1. The mean and variance parameters here are simply examples, the important aspect is the correlation structure imposed. In this paper, we require $\rho_1 = \rho_2 = \rho_3$, but this would not be necessary in general.

1.2.1 Generative Model

The data generating process for the purpose of our inference is

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim \mathsf{No}\left(\mu, \Sigma\right)$$

$$\phi_i \mid Y_i = \pi_{SB}(Y_i)$$

$$X_i \mid \phi_i \sim \sum_{j=1}^{k} \phi_i^j g(x \mid \theta_j).$$

Here $g(x \mid \theta_j)$ is a known distribution. To facilitate inference, we add latent variables Z_i^j that represent which component of the mixture from which X_i^j was drawn. This means that Z_i^j is a categorical variable taking values from 1 to k with probabilities ϕ_i. With this latent variable the DGP becomes

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim \mathsf{No}\left(\mu, \Sigma\right)$$

$$\phi_i \mid Y_i = \pi_{SB}(Y_i)$$

$$Z_i^j \mid \phi_i \sim \mathsf{Cat}(1, \psi_i)$$

$$X_i^j \mid Z_i \sim g(x \mid \theta_{Z_i^j}).$$

The important aspect of the model is that Y_1 and Y_2 are not independent. We will develop two models, with different assumptions on Σ. In the first, ρ will be a tuning parameter that determines how related the probability vectors are assumed to be. In the second, the value of ρ will be learned within the model.

1.2.2 Posterior Inference with fixed ρ

Suppose we fix the value of ρ at some value in $[-1, 1]$ and use the structure for Σ given in Equation 1.4. The goal is to sample from the posterior distribution of ϕ given X. As we have written it now, this involves two full conditions: $Y \mid Z, X$ and $Z \mid \phi, X$. The update for $Z \mid \phi, X$ is standard, with the details given in the appendix. The update $Y \mid Z, X$ is not conjugate, but can be make conjugate by means of data augmentation.

1.2.3 Pólya-gamma data augmentation

The Pólya-gamma family of distributions is characterized by two parameters $\mathsf{PG}(b, c)$, which are defined here. The distribution was first introduced by Polson et al. (2013). A variable $\omega \sim \mathsf{PG}(b, 0), b > 0$ if it is equal in distribution to an infinite sum of gamma random variables:

$$\omega \stackrel{d}{=} \frac{1}{2\pi^2} \sum_{k=1}^{\infty} \frac{g_k}{(k - 1/2)^2}, \tag{1.5}$$

where $g_k \sim \mathsf{Ga}(b, 1)$ are independent random variables. The general distribution is defined by exponentially tilting of $\mathsf{PG}(b, 0)$. The density is given by

$$p(\omega \mid b, c) = \frac{\exp\left(-\frac{c^2}{2}\omega\right) p(\omega \mid b, 0)}{\mathsf{E}_\omega\left\{\exp\left(-\frac{c^2}{2}\omega\right)\right\}}, \tag{1.6}$$

with $p(\omega \mid b, 0)$ representing the density of a $\mathsf{PG}(b, 0)$ random variable. The expectation is with respect to a $\mathsf{PG}(b, 0)$ random variable. Using this distribution we will be able to construct a Gibbs' sampler to converge to the posterior.

1.2.4 Posterior Inference with variable ρ

The previous method suffers from the fact that we need to choose the covariance ρ. As with any such tuning parameters, we must select appropriate values. A poor

choice will cause the method to perform very poorly, so the question arises how we should pick ρ. We see that learning ρ is the same as learning a covariance. To do this, we make use of an algorithm developed in Liu and Daniels (2006).

The algorithm from Liu and Daniels (2006) is designed for learning correlation matrices in the context of multivariate probit models and multivariate regression models with a common correlation matrix across groups. To adapt the algorithms to our context, we simply set the design matrix $\mathbf{X}$ to be a matrix of zeros. With this normalization, we can describe the algorithm in this context. Let R denote a correlation matrix. In our context, suppose we have two probability vectors ψ_1 and ψ_2. We have

$$R = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix},$$

where $\rho = \mathsf{Cor}\left(Y_{1,i}, Y_{2,i}\right)$. In other words, the value of ρ is the correlation between the components of Y_1 and Y_2. Since ψ_1^i and ψ_2^i are calculated deterministically from Y_1 and Y_2, we are inducing a stochastic relationship between these components.

Let $R^{(n)}$ and $Y_i^{(n)}$ be the current values of R and Y_i for $i = 1, 2$ Let $\mathbf{Y}$ denote the matrix with whose columns are Y_1 and Y_2. Proceed as follows:

1. Calculate
$$D = \begin{pmatrix} \sum_{j=1}^{k} \left(Y_1^{j(n)}\right)^2 & 0 \\ 0 & \sum_{j=1}^{k} \left(Y_1^{j(n)}\right)^2 \end{pmatrix}.$$

2. Define $\mathbf{Y}^*$ as the matrix given by $\mathbf{Y}^* = \mathbf{Y}D$.

3. Let $S = \mathbf{Y}^{*\top}\mathbf{Y}^*$ and sample $\Sigma \sim \mathsf{InvWi}(k, S^{-1})$. Extract the correlation matrix $R^* = \operatorname{diag}(\Sigma)^{-1/2}\Sigma \operatorname{diag}(\Sigma)^{-1/2}$.

4. Calculate $\alpha = \min\left\{1, \exp\left(\frac{T+1}{2}\left(\log|R^*| - \log|R^{(n)}|\right)\right)\right\}$, the Metropolis-Hasting acceptance probability. Here $T = 2$ is the number of Y vectors we have.

5. Accept R^* with probability α, in which case we set $R^{(n+1)} = R^*$. Otherwise $R^{(n+1)} = R^{(n)}$.

The method generalizes to more than two vectors Y_1, Y_2. This method assumes a uniform prior for R:

$$\pi(R) \propto \mathbf{1}\left\{R_{jk} : R_{jk} = 1 \text{ if } j = k \text{ and } |R_{jk}| < 1 \text{ if } j \neq k \text{ and } R \text{ is positive definite.}\right\}.$$

1.2.5 Prior Selection

We need to select values for μ and Σ, which induce the prior on ψ_1 and ψ_2. Selecting a suitable prior in this circumstance is very difficult. Note that ψ is stochastically ordered; the components are not exchangeable. To determine the prior, consider the symmetric Dirichlet distribution with concentration parameter $\alpha = 1$, which is uniform over all points in the probability simplex.

We want the joint prior distribution of ψ to approximate the symmetric distribution over the simplex. An approximation can be achieved by exploiting the fact that the stick-breaking map is invertible. This invertibility suggests the following procedure to calculate μ and Σ:

1. Sample N (say 10,000) independent draws from $\eta_i \sim \mathsf{Dir}(\alpha = 1)$ for $i = 1, \ldots, N$.

2. Apply the inverse stick-breaking map to get vectors $X_i = \pi_{SB}^{-1}(\eta_i)$ for $i = 1, \ldots, N$.

3. For each component j of X_i^j estimate the mean and variance of a normal with the MLE estimators $\hat{\mu}_j = \bar{X}^j$ and $\hat{\sigma}_j^2 = \hat{\sigma}_{X^j}^2$.

If $k = 4$, the resulting densities are shown in Figure 1.1. The four dimension probability vectors are determined by three components. The resulting densities are approximately normal, but clearly exhibit a slight left tail. The Shapiro-Wilk test

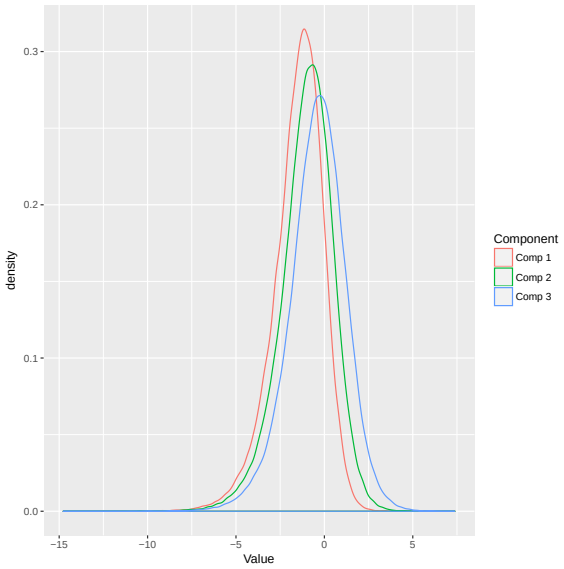

FIGURE 1.1: Densities for X^j resulting from pull-back of symmetric Dirichlet through stick-breaking map.

rejects the normality of all components. To get our prior, we will approximate these densities with normal densities using the MLE to estimate the mean and variance. The results of the approximation are shown in Figure 1.2. The density that we are trying to approximate is $\mathsf{Be}(1, 3)$, which is shown by the black line. The various colored lines are the results of stick-breaking draws from our the normals with the parameters determined as above. The approximation is probably adequate for our purposes.

1.3 Tests on Synthetic Data

Suppose we have two "similar" probability vectors

$$\psi_1 = (0.25, 0.25, 0.25, 0.25),$$

$$\psi_2 = (0.23, 0.27, 0.24, 0.26).$$

Fix the kernels to be $g(x \mid \theta_j)$ to be $\mathsf{No}(1 + 3(j - 1), 1)$ (normals with means 1,4,7,10 and variance 1). Draw a sample of size 500 from f_1 (a "large" sample) and a sample of size 100 from f_2. The density estimates for a these samples are shown in Figure

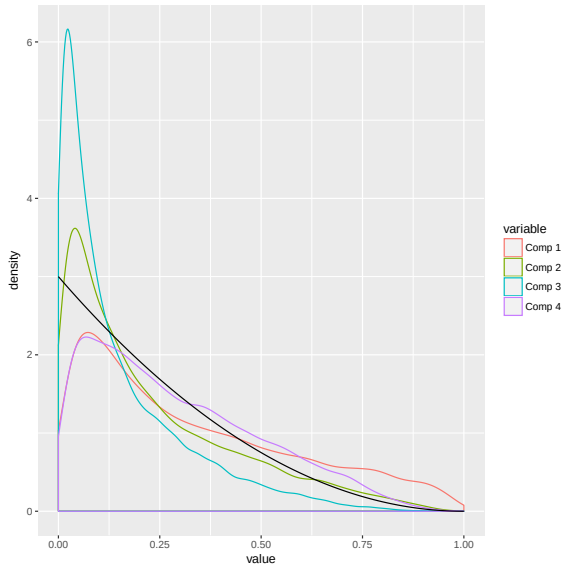

FIGURE 1.2: Marginal densities after stick breaking. The black line represents the density of $\mathsf{Be}(1, 3)$. .

1.3. Note that the density estimate for f_1 shows four distinct modes in approximately the correct places. The modes should all be equal, however, which we do not observe. The density estimate for f_2 only has two modes. This shows that the sample is small enough the it is not trivial to identify the features of the sample. In particular, it is not at all apparent that there are four components in the mixture distribution. Thus these sample sizes are not so large that inference is trivial.

In Figure 1.4 we look at our estimates for $\psi_1^1 = 0.25$ with four different covariance levels ($\rho = 0, .5, 0.9, 0.99$). The figure shows that for the large sample, the covariance ρ does not greatly affect posterior inference. The posterior modes are similar to the MLE, with some covariance levels slightly closer to the true value. The results are reassuring, as the performance here is comparable to the MLE.

In Figure 1.5 we have various estimates for ψ_1^4. Note that the posterior mode for $\rho = 0.99$ is the best estimate. The MLE outperforms some estimates with small ρ. This may result from the fact that our estimates, which rely on stick breaking, are stochastically ordered. The stochastic order might result allow the first components

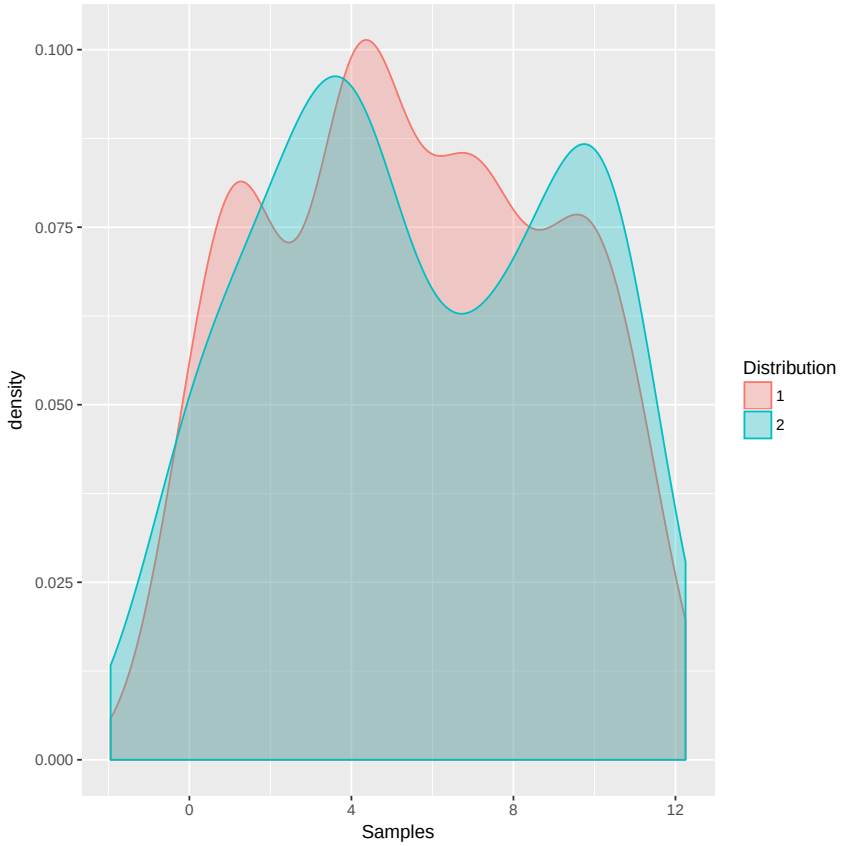

FIGURE 1.3: Kernel density estimates for the samples from f_1 and f_2.

of ψ_1 to be accurately estimated, but adversely affect the remaining estimates. We have no strong evidence of that.

The more interesting question is how well the method can estimate ψ_2, where we have significantly less data. The estimates for $\psi_2^1 = 0.23$ are shown in Figure 1.6. Both the MLE and the posterior mode give are close to the true value. Note that the posterior variance is much higher for ψ_2, which is expected because we are basing our inference on a larger sample. The estimates for ψ_2^2 are shown in Figure 1.7. Here our method begins to show some advantages. The posterior mode is closer to the true value for most values of ρ. In addition, when the covariance is $\rho = 0.99$, the posterior variance is noticeably smaller. This suggests that we are indeed able to share information from the large sample to improve the accuracy of our estimates

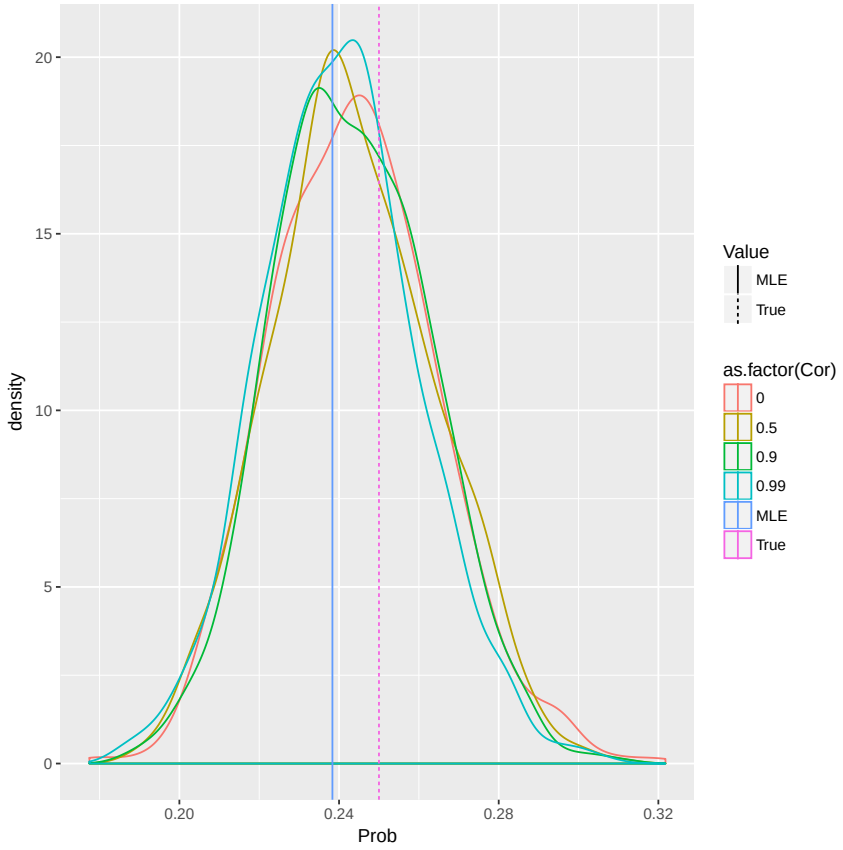

FIGURE 1.4: Posterior estimates for ψ_1^1 with various ρ.

and reduce the variance. The estimates for ψ_2^3 and ψ_2^4 are shown in Figure 1.8. Both figures suggest that by using a large ρ we are able to improve both the precision and accuracy of our estimates.

1.4 Conclusion

In this chapter we discuss a Bayesian model for spike train data, which we take as a motivation for developing a method to share information and learn multiple probability vectors jointly. Our method uses stick breaking and Pólya-gamma data augmentation to sample correlated probability vectors. We test our method on a mixture two mixture distributions with known kernels and find that by inducing correlation between the vectors we are able to improve our estimates when the sample

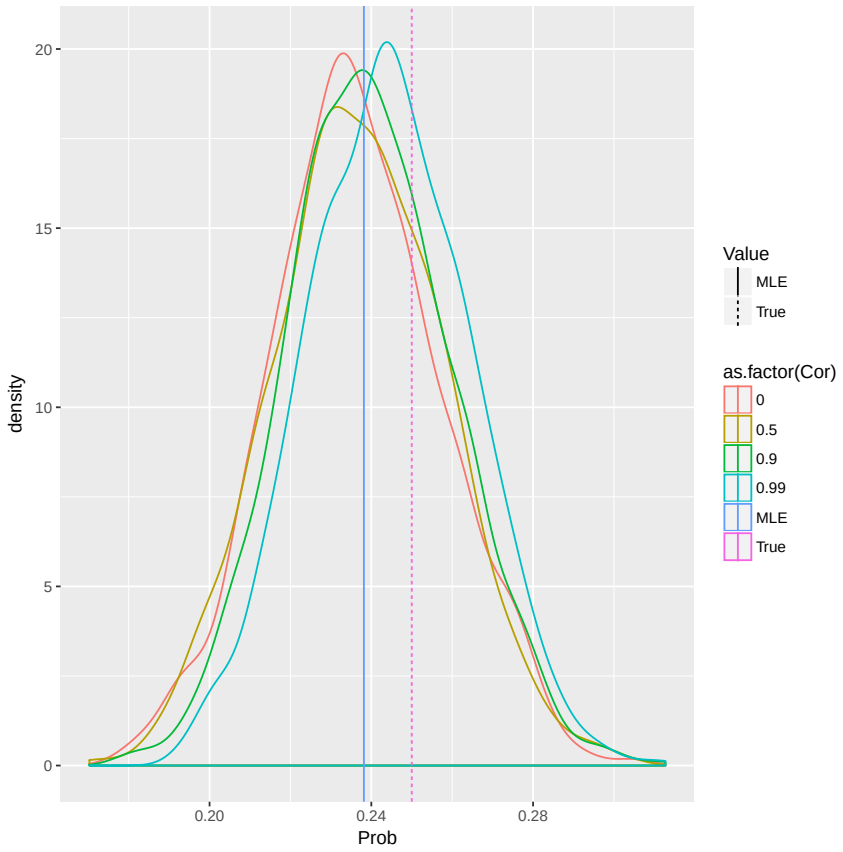

FIGURE 1.5: Posterior estimates for ψ_1^4 with various ρ.

size is small.

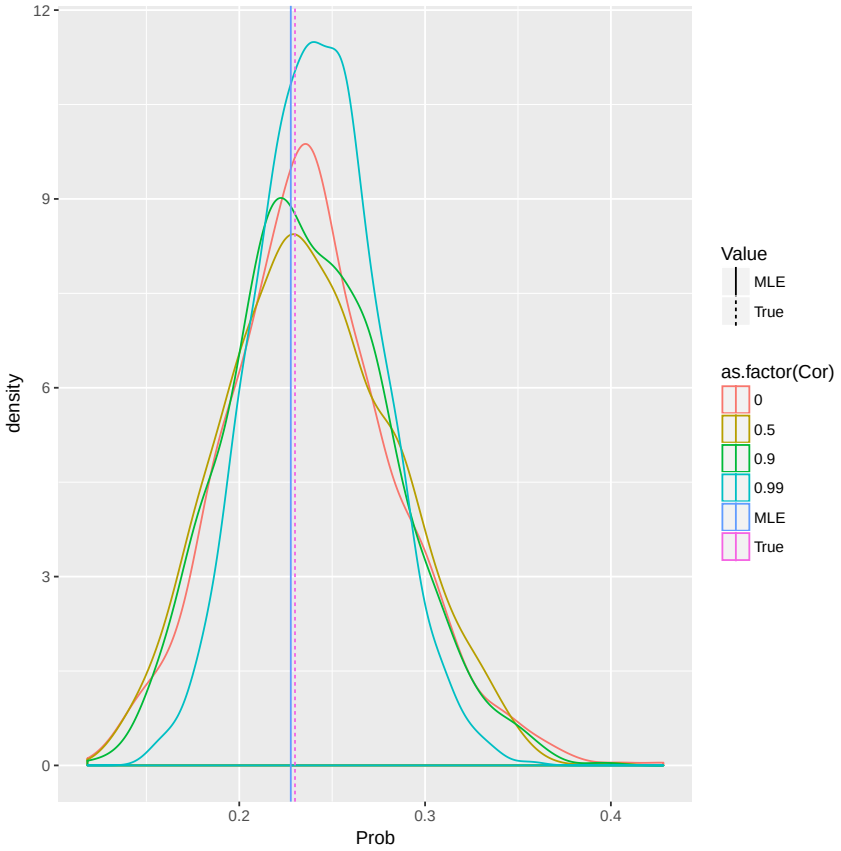

FIGURE 1.6: Posterior estimates for ψ_2^1 with various ρ.

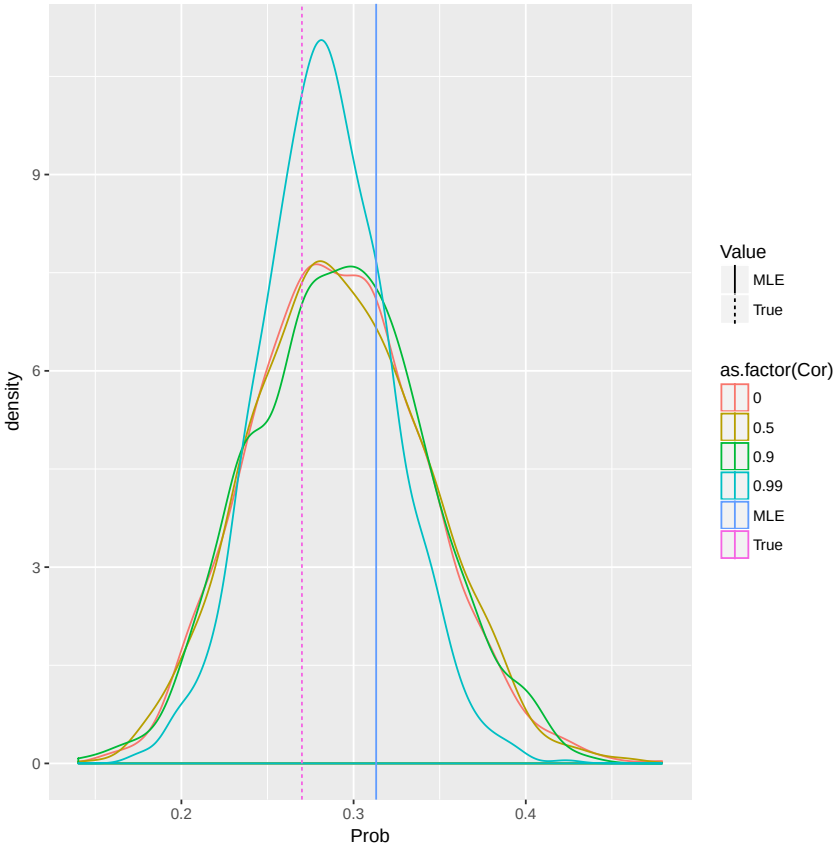

FIGURE 1.7: Posterior estimates for ψ_2^2 with various ρ.

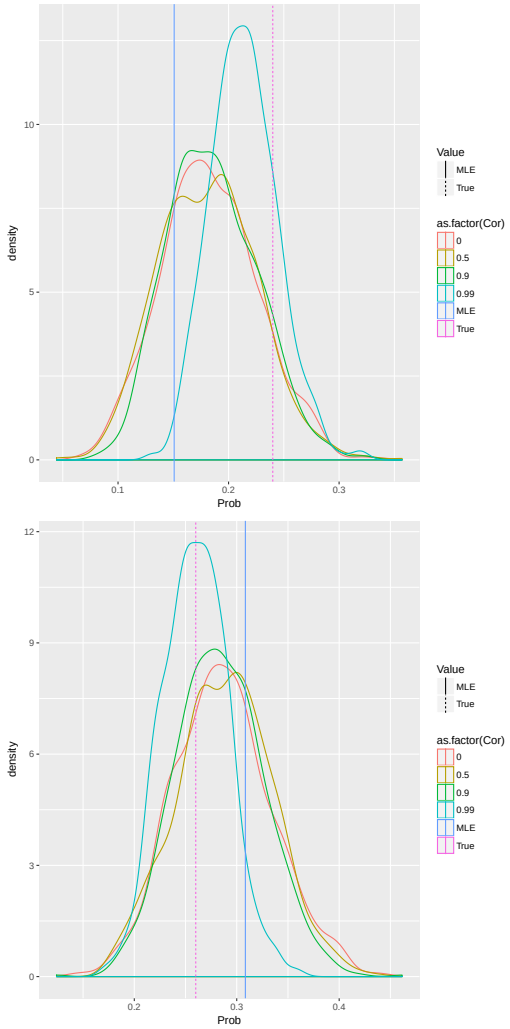

FIGURE 1.8: Posterior estimates for ψ_2^3 and ψ_2^4 with various ρ.

2

Bayesian Analysis of Spectrum Auctions

2.1 Introduction to Spectrum Auctions

The Federal Communications Commission (FCC) is an independent federal agency tasked with regulating various types of communications, including radio and television. In 1994 the FCC began to allocate commercial spectrum licenses using auctions. The spectrum sold has many uses, including television broadcasting, cellular communication, and radio. The rise of advanced cellular phones capable of transmitting data created a surge in demand for spectrum suitable for such transmission.

Due to the physical properties of spectrum near 700 MHz, licenses containing spectrum in this range are especially valuable. In 2008, the FCC help an auction where it sold these valuable licenses for regions throughout the country. Some of these licenses sold for millions of dollars. In fact, the winning bids in the auction totaled $18 billion. Our goal is to understand what factors contribute to the value of these licenses.

2.1.1 Previous work

As mentioned above, these licenses are incredibly valuable. Huge companies like Verizon Wireless and AT&T invest massive amounts of money to win these licenses, which support their cellular networks. Naturally, this has driven much research into these auctions. For a good overview of the existing literature, see Connolly et al. (2017).

2.2 Theoretical Auction Model

Suppose that there is a license L being auction in an ascending multi-round auction (for now we assume that there is a single license or that bidders' strategies in all licenses are completely independent). There is a set of bidders $\mathcal{B}$ with bidders indexed as $i = 1, \ldots, n \in \mathcal{B}$. The rounds are discrete and any bidder can bid in any round. Bidders must bid in fixed bid increments; a new bid must exceed an existing bid by an amount T. The auction ends when there are no new bids in a round. If two bidders bid the same amount, the winner is decided uniformly at random. Once the auction concludes, the winning bidder pays the highest bid.

We denote the valuation of bidder i as V_i. The valuation is the maximum amount that a bidder would be willing to pay for L. If the price of L exceeds V_i, the bidder would prefer to not buy the license, rather than pay any amount greater than V_i. For concreteness, we assume that if the price is exactly V_i, the bidder will place a bid.

The bid of bidder i in round k will be denoted $b_{i,k}$.

2.2.1 Analysis of Auction Model

As with many problems in auction theory, we are interested in recovering the valuations V_i after observing the results of the auction. The structure of the auction will make this problem very tractable. Consider a bidder i who has placed a bid $b_{i,k}$,

which is exceed in round $k+1$ by a bid $b_{j,k+1} = b_{i,k} + T$ for $i \neq j$. When is it rational for bidder i to place a bid of $b_{i,k} + 2T$ in round $k + 2$? Clearly, this is rational if and only if $b_{i,k} + 2T \leqslant V_i$. So if the bidder does not bid, we can conclude that the bidder must have realized that $b_{i,k} + 2T > V_i$, in which case bidding could result in negative utility. Together with the fact that the bidder placed $b_{i,k}$, we can conclude $b_{i,k} \leqslant V_i < b_{i,k} + 2T$. Depending on the size of T, the bound we achieve may be very tight.

The analysis of the auction above is very sound if there were only one license. Once we step back and acknowledge that a license is being sold within a larger auction there are a few issues that become immediatley obvious. The first is that there are multiple licenses being sold in the auction, and a bidder can substitute between various licenses. The problem is that if a bidder decides not to bid, it need not be the case that $b_{i,k} + 2T > V_i$, it could instead be the case that the bidder could achieve higher overall utility by giving up the current license and pursuing a different one.

2.3 The Data

In this section we give some summary statistics for results of auction 73. General statistics are given in Table 2.1. Small bidders are those who receive a discount (of either 15% or 25%). For instance, if a small bidder places a bid of $1000, they would only pay the FCC $750. These discounts were introduced to encourage new companies to enter the market. The average winning bid is deceptive because the distribution of winning bids is highly skewed. A histogram of the winning bids on a logarithm scale is given in figure 2.1.

Each license covers a certain area of the United States. Some licenses are national licenses that cover the entire country, while others are regions licenses that may cover a single city. The regions used in auction 73 are generally Economic Areas (BEAs)

Table 2.1: Summary Statistics for Auction 73

Number of Licenses Sold	1,090
Number of Bidders	207
Number of Winners	101
Number of Small Bidders	115
Average Winning Bid	$17,825,343

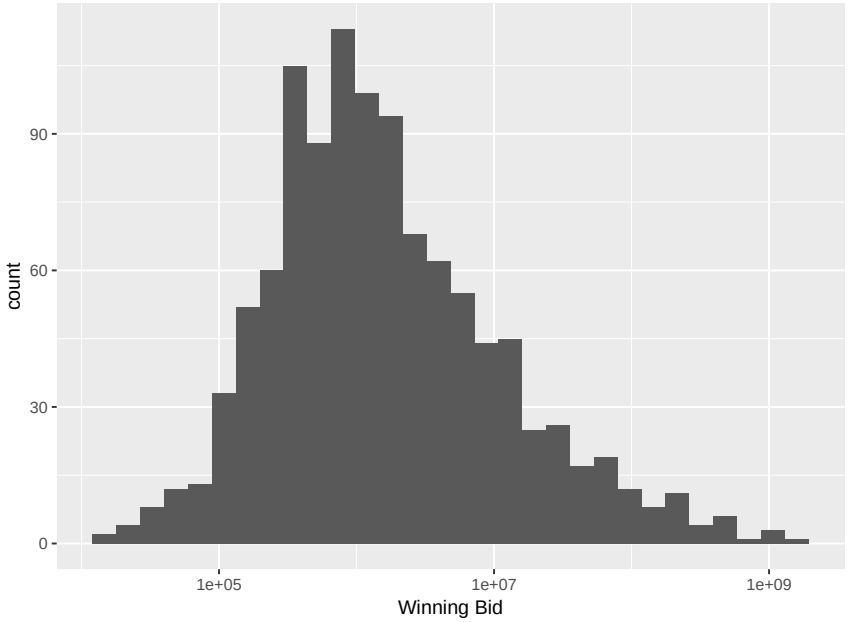

FIGURE 2.1: A histogram of the winning bids on a log scale.

and Cellular Market Areas (CMAs). The CMA regions are shown in Figure 2.2.

2.3.1 Important Variables

In this section we give a brief overview of the important variables in the data and their meaning. The two central variables are `BID_AMNT` and `NET_BID_AMNT`. The variable `BID_AMNT` tells us the size of the bid placed by each bidder in each round. The variable `NET_BID_AMNT` is the amount that the bidder would have to bid after any discounts. We always have

$$\texttt{NET\_BID\_AMNT} = \delta\left(\texttt{BID\_AMNT}\right)$$

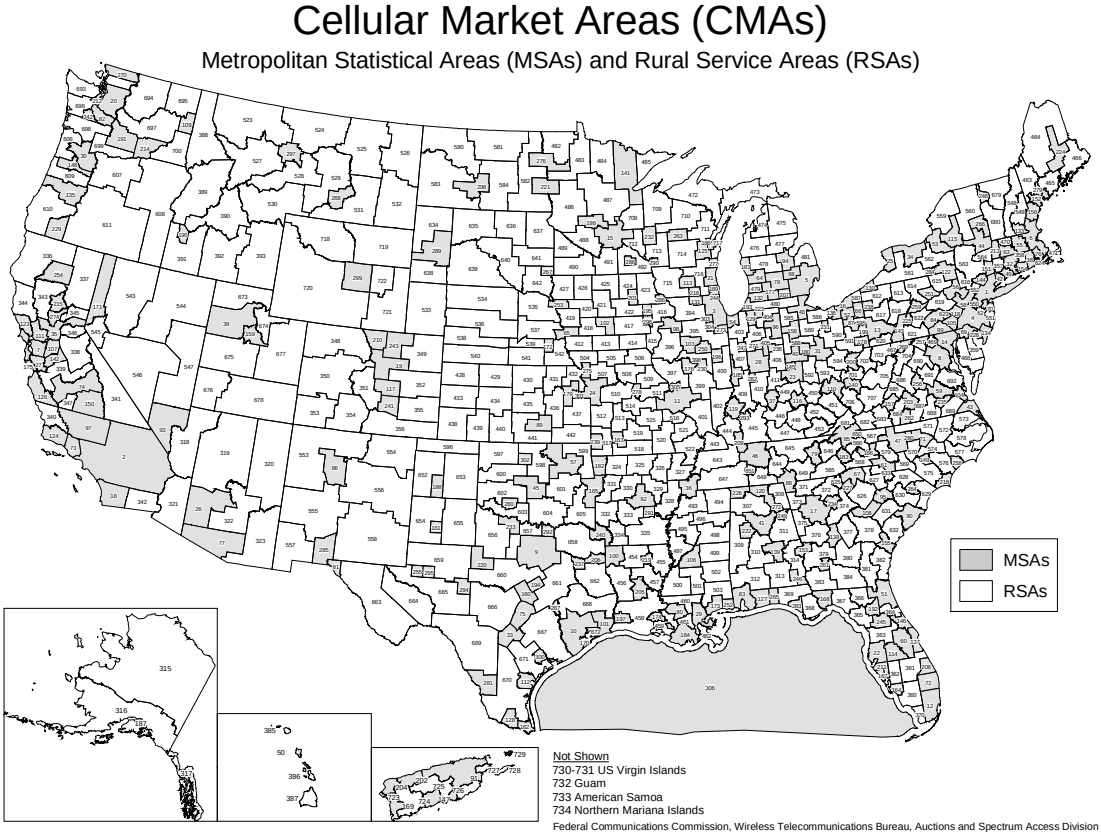

FIGURE 2.2: Map of Cellular Market Areas.

for some $\delta \in \{1, .85, .75\}$. Another important variable is `MAX_ELIG`. In order to participate in an auction, firms must make a deposit with the FCC. A histogram of these deposits (on a log scale) are shown in Figure 2.3. This deposit determines how much the firm is allowed to bid in an auction. As the firm places larger and larger bids, the eligibility of the firm is reduced in relation to the bids. If the eligibility reaches zero, a firm will not be allowed to place new bids.The system is designed to prevent firms from placing huge bids that they will be unable to actually pay. The variable is important because it tells us about the size of a firm. Firms that have large initial eligibility are those that are determined to acquire spectrum. A small firm cannot deposit as much money as Verizon or AT&T. In this way, the variable tells us about firms size and wealth. We will generally work with the firm's eligibility in the first

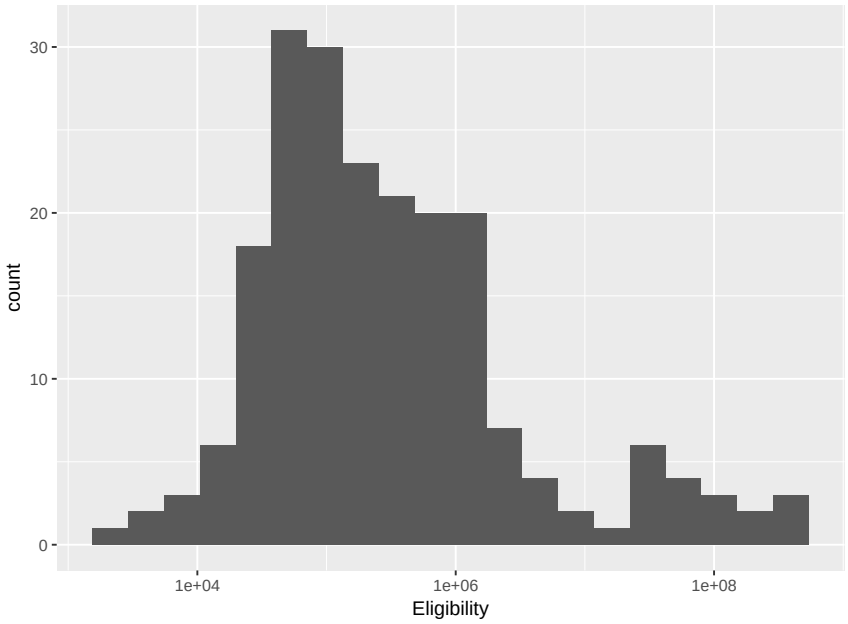

FIGURE 2.3: Histogram of initial deposits to the FCC made by bidders before Auction 73.

round, which is the most informative about the firm's size. Looking at Figure 2.3, we can see the separation between small firms and large firms.

For each license area, we have `POPULATION`, the population of each license area. Population is important because the value of a license should be closely related to the population. Firms purchase spectrum to provide service to those living or traveling through the region covered by a license. An license like the Gulf of Mexico (CMA306) will be valued differently than a license that covers New York City (CMA001). There are many more potential customers in New York than there are in the middle of the Gulf. Our regressions will include both population and population squared, scaled by a factor of 1000 for interpretability.

In our regressions we also control for `BANDWIDTH`, which is the size of the license. For instance, if the license covers from 740MHz to 750MHz, the bandwidth is 10MHz. There are standard license sizes, as shown in Table 2.2. Most licenses sold have a bandwidth of 12MHz.

Table 2.2: Breakdown of License Bandwidth in Auction 73

Bandwidth (MHz)	Number of Licenses
6	176
10	1
12	902
22	11

2.3.2 Auction Example

In this section we look at the progression of bids for New York (CMA001) and Chicago (CMA003). In Figure 2.4 we see the bids placed in the first 10 rounds of the auction. The size of the circle reflects the size of the firm (as indicated by `MAX_ELIG`). A small amount of random noise has been added to visually separate the points. In reality, all firms placed the exact same bid. The fixed bid increments were implemented after firms used trailing bids to signal one another in early auctions. To discourage such collusion, the FCC fixed the bid increment. The bid increment is actually set to increase by 10% over the previous bid, which results in the smooth curve through the points.

For both licenses, we see many firms competing in the initial rounds, including small firms. The small firms are quickly bid out. In the later rounds (see Figure 2.5), we see a few larger firms remain. The important feature is the alternating pattern that emerges. We generally stop seeing multiple firms bidding in one round. Instead, firms take turns raising their bids. This pattern is very appealing because it enhances our analysis of such auctions. Once a firm stops bidding (if it is in such a step pattern), we have strong evidence that it has reached its valuation. In short, the bidding strategies we observe in the data are encouraging based on our theoretical analysis.

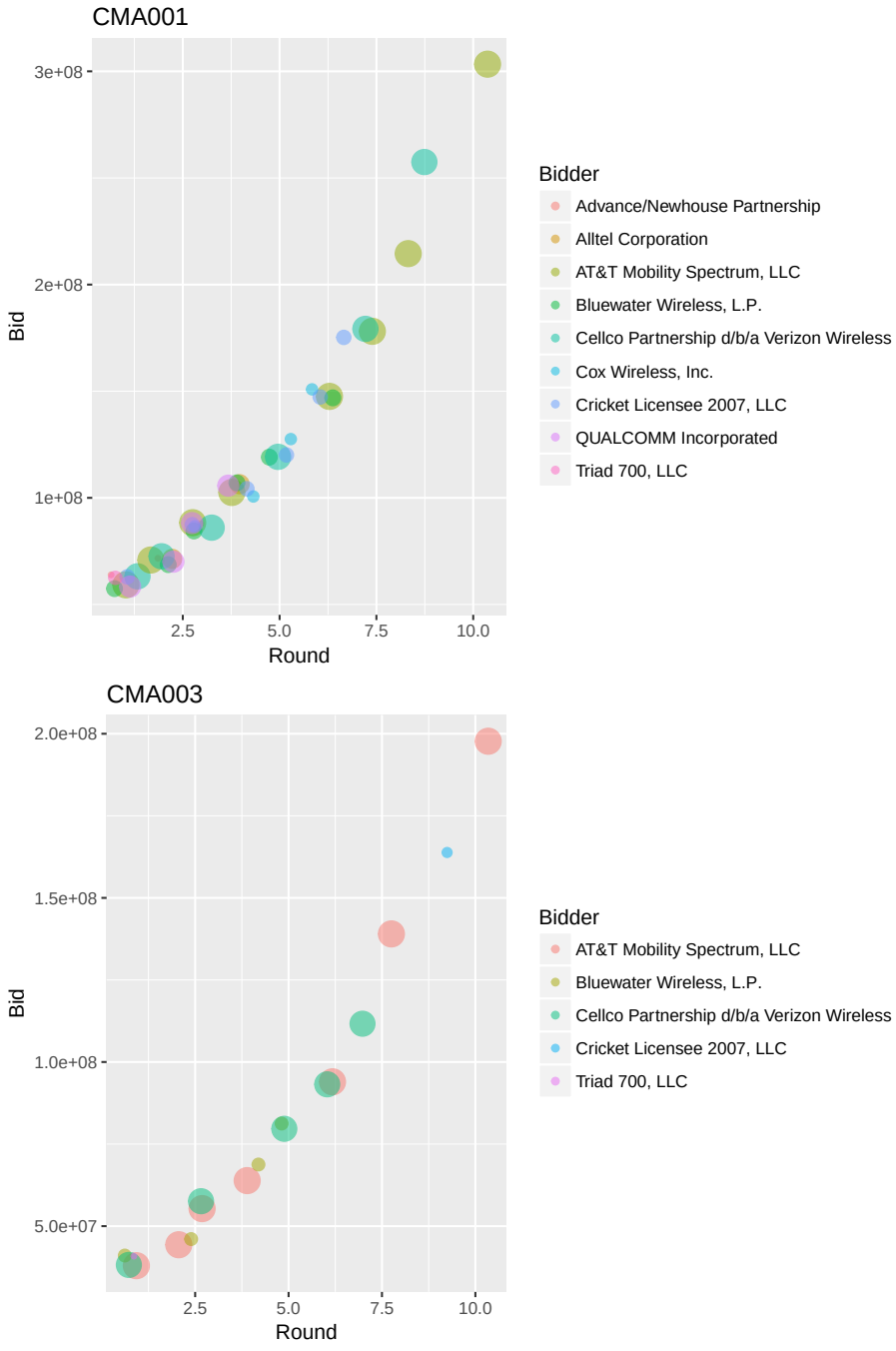

FIGURE 2.4: Bids in the first ten rounds for the licenses in New York (CMA001) and Chicago (CMA003). The size of the points reflects the size of the deposit made by the firm.

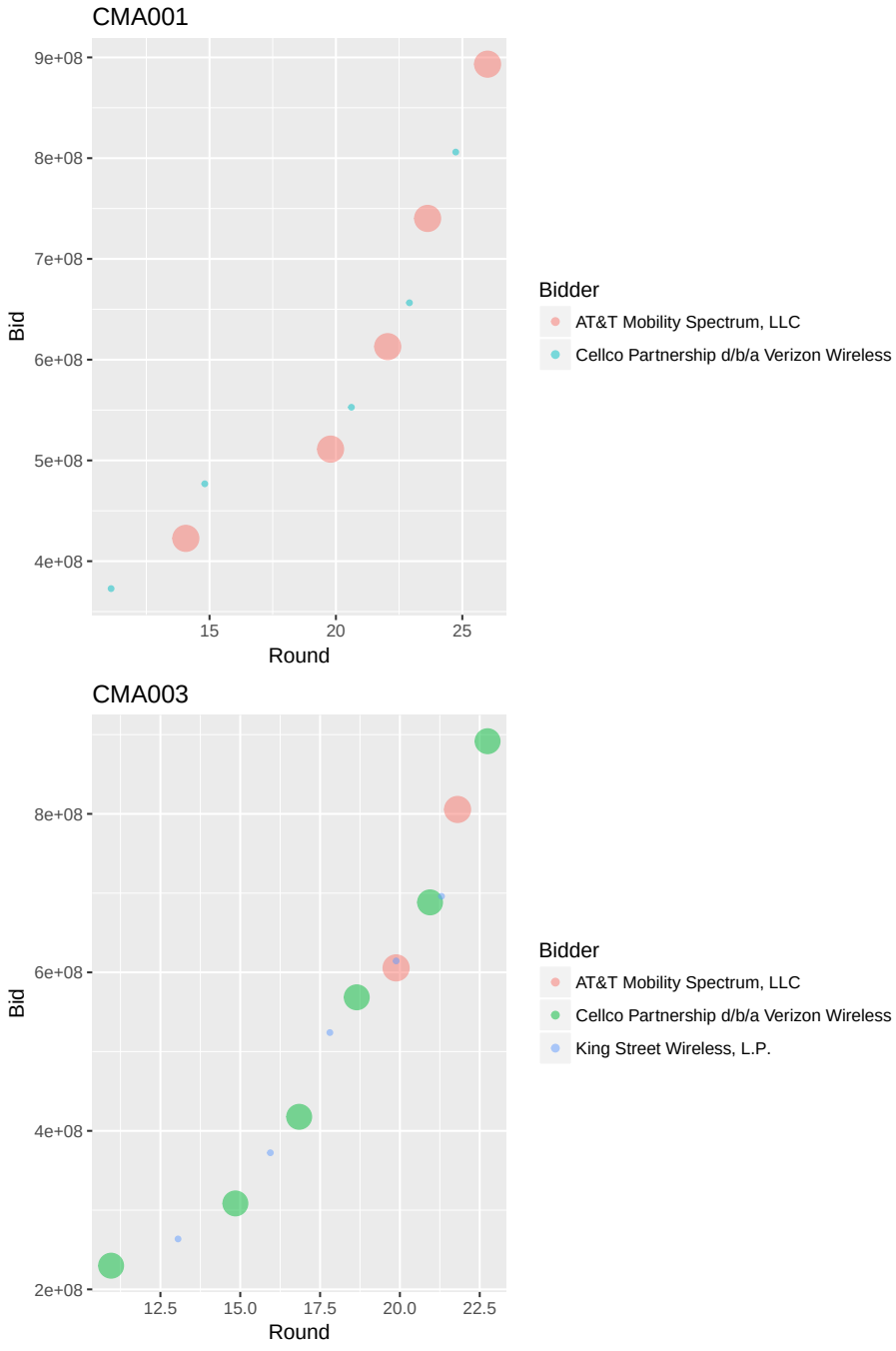

FIGURE 2.5: Bids in the later rounds for the licenses in New York (CMA001) and Chicago (CMA003). The size of the points reflects the size of the deposit made by the firm.

2.4 Statistical Model

Given the analysis above, we are able to conclude with fairly high confidence that the true valuation is in some (possibly unbounded) interval. This is a classical example of a censored data problem, where we do not observe the true value, but rather a rounded value or bound. Bayesian are extremely well suited to deal with this problem.

2.4.1 The Censoring Problem

The data collected in auctions are censored. We observe the bids placed by the bidders, but not their valuations. These valuations are what are typically of interest, as we hope to explain them in terms of covariates. The bids have seemingly little connection to underlying valuations. For example, consider a dollar auction, where each person bids on a one dollar bill. If each person can bid successively on a one dollar bill, what will the result be? The bill will probably sell for \$.99 or \$1, but the bids we observe are not reflective of the valuations, which should all be equal to \$1. The bids are strategic choices, somewhat independent of the underlying economic principles captured by the valuation.

Ordinary least squares estimation falls apart in the presence of censoring. The estimates are biased and inconsistent. To address this issue, some articles in the literature turn to the Tobit model. The Tobit is estimated by maximum likelihood, in which the contribution of the censored values is a assumed to come from a normal distribution. For instance, if $x_i \geqslant 200$ and we observe 200, then the likelihood contribution is $1 - F(200)$, where F is the normal CDF with appropriate mean and variance. Using the Tobit corrects for the problems created by censoring. Most works that use the Tobit model only treat the winning bid as censored and ignore the non-winning bids. In this work, we treat find regions as above for *all* bidders.

2.4.2 Bayesian Regression

To address the censoring problem, we will use a Bayesian linear regression. To fit this regression, we will use a Gibbs' sampler where the bidder valuations V_i are imputed in each sweep. As one of our goals is to determine the effects of ignoring the censoring problem (say, by using OLS), we will use a Normal-Normal model, with the following DGP for a license L:

$$\beta_L \sim \mathsf{No}(\mu_0, \Sigma_0)$$

$$\mathbf{V}_L \mid \mathbf{X}, \beta \sim \mathsf{No}(\mathbf{X}_L \beta_L, \Sigma_L)$$

where $\mathbf{X}_L$ is a matrix of covariates and

$$\mathbf{V}_L = \begin{pmatrix} V_1 \\ \vdots \\ V_N \end{pmatrix}$$

is the vector of valuations for license L. First, not that we do not observe V_i; we observe an interval, which we believe, based on our theoretical reasoning about the structure of the auction, should contain V_i. We can set up a Gibbs' sampler where we impute the missing values in $\mathbf{V}_L$. The design matrix $\mathbf{X}_L$ can be thought of as containing two different types of variables. One one hand, it contains license characteristics, which are the same for all bidders. This would include things like bandwidth of the license or the population of the license area. Of course, if we only have a single license the coefficients cannot be estimated. There are also firm specific variables, such as the eligibility.

An important question is what structure we want to impose on Σ_L. We will assume that different companies determine their valuations *independently*, so Σ_L is a diagonal matrix. If we are imitating the assumptions of an OLS or Tobit model, we further impose $\Sigma_L = \sigma^2 I_N$ for some variance σ^2.

The next question is how to aggregate across licenses. The simplest method would be to combine the licenses independently. This again matches the assumptions that would be used in most standard regression models. So if $\mathbf{V}$ is the set of all valuations for all bidders across all licenses, then our generative model is

$$\beta \sim \mathsf{No}(\mu_0, \Sigma_0)$$
$$\mathbf{V} \mid \mathbf{X}, \beta \sim \mathsf{No}(\mathbf{X}\beta, \sigma^2 \mathbf{I}_N).$$

Sampling from posterior of this regression model is very simple with a Gibbs' sampler. We update

1. Given the current value of β and σ^2, sample V_i from an appropriatly truncated normal.

2. With the imputed values of $\mathbf{V}$, update β and σ^2 using the traditional regular formulas.

This process is shown in Figure 2.6. By imputing at every iteration, we more accurately incorporate our uncertainty about the value of V_i. This is in contrast to regressing on a small number (5 or 10) of imputed data sets, which reduces the amount of uncertainty in our estimates.

2.4.3 Prior Selection

As with all Bayesian models, we must select a prior distribution for the parameter β. We will use the unit information prior. For a regression model, this centers the prior at the OLS estimate for β, which we denote $\hat{\beta}_{OLS}$. The variance is set to be the variance we would have with one observation. So if $\mathbf{V}_{obs}$ are the observed bids, then the prior parameters are

$$\begin{aligned}\mu_0 &= \hat{\beta}_{OLS} \\ &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{V}_{obs}\end{aligned}$$

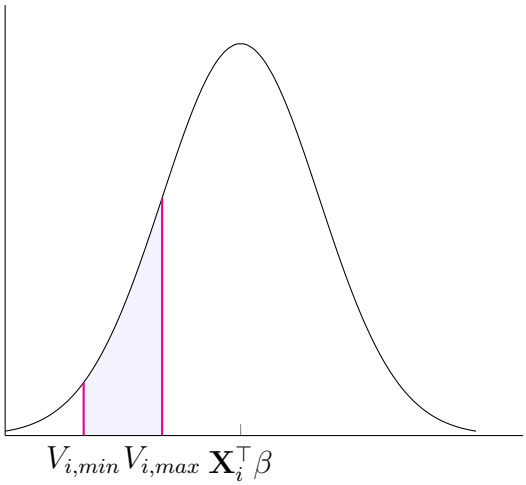

FIGURE 2.6: At each iteration, we impute V_i from a truncated normal indicated by the shaded region.

and

$$\Sigma_0 = n\sigma^2_{OLS}(\mathbf{X}^\top\mathbf{X})^{-1}$$

where

$$\sigma^2_{OLS} = \frac{(\mathbf{V}_{obs} - \mathbf{X}\hat{\beta}_{OLS})^\top(\mathbf{V}_{obs} - \mathbf{X}\hat{\beta}_{OLS})}{n-p)},$$

with p denoting the number of parameters. If our desire is to demonstrate that OLS suffers from the problems we expect, then this prior is very suitable. By centering the prior at the OLS estimate (which is also the MLE), we are saying that our prior beliefs are that OLS is correct. This means that any difference that we observe between the prior and the posterior are coming from the data, not from our choice of prior. There are those who may raise objections to the use of the unit information prior because the prior parameters are functions of the data, but we will proceed without becoming bogged down in this issue.

2.4.4 Comparison of Assumptions

The ordinary least squares regression model is built on five assumptions.

Assumption 1. *The observed data is linear:* $\mathbf{Y} = \mathbf{X}\beta + \epsilon$.

Assumption 2. *The error term is mean independent of* $\mathbf{X}$*:* $\mathsf{E}[\epsilon \mid X] = \mathbf{0}$.

Assumption 3. *The matrix* $\mathbf{X}^\top\mathbf{X}$ *is invertible.*

Assumption 4. *Homoskedasticity:* $\mathsf{V}[\epsilon \mid X] = \sigma^2\mathbf{I}_n$.

Assumption 5. *Normally distributed errors:* $\epsilon \mid X \sim \mathsf{No}(\mathbf{0}, \sigma^2\mathbf{I}_n)$.

Assumptions 1-4 are the basis for the Gauss-Markov theorem, which shows that OLS is efficient. Assumption 5 gives us the finite sample distribution for $\hat{\beta}_{OLS}$. These assumptions are essentially identical to the assumptions that form the basis of our Bayesian model. Since we assume that $\mathbf{V} \mid \mathbf{X}, \beta \sim \mathsf{No}(\mathbf{X}\beta, \sigma^2\mathbf{I}_n)$, we have that $\mathbf{V} = \mathbf{X}\beta + \epsilon$, where $\epsilon \mid \mathbf{X}, \beta \sim \mathsf{No}(\mathbf{0}, \sigma^2\mathbf{I}_n)$. This gives us Assumptions 1, 2, 4, and 5. Assumption 4 is a technical assumption, without which we could not compute the posterior. This means that our model makes the same assumptions as OLS. The main difference is that in the Bayesian model we consider the distribution of ϵ conditional on $\mathbf{X}$ *and* β. This is because in a Bayesian model, β is a random variable and not a fixed parameter like in OLS.

2.4.5 Bayesian Regression Incorporating Spatial Considerations

We made many simplifying assumptions in the model above to mimic the assumptions made for OLS. For an economic perspective, it may be desirable to have a more complex model. We will now extend the model to incorporate geographic considerations.

Suppose that a wireless provider has a contract with a customer in New York. The provider wants to ensure that the customer has good coverage in most places they travel. To do this, the provider will need a license that covers New York. However, it is highly likely that the customer may travel to nearby cities or suburbs. To cover

the customer in these areas, the provider will need to own licenses that are adjacent to New York as well. In general, it makes sense for a provider to own a set of licenses that cover a contiguous area.

To incorporate this into a model, suppose that a provider i bids on a set of licenses $L_{i,1}, \ldots, L_{i,k}$. Consider a matrix T_i where the $t_{j,k} = 1$ if the region covered by $L_{i,j}$ touches the region covered by $L_{i,k}$. Set $t_{j,j} = 0$. So if $\mathbf{V}_i$ is the vector of valuations for bidder i, then we model

$$\mathbf{V}_i \mid X_i, \beta \sim \mathsf{No}(\mathbf{X}_i\beta, \sigma^2(\mathbf{I}_n + \tau T_i)).$$

This means that for any two licenses, $L_{i,j}$ and $L_{i,k}$, either the valuations are independent or they are correlated with correlation τ. So when the bidder draws the valuations, licenses that do not touch are independent, but licenses that do touch are correlated with correlation τ. The goal is to learn τ as well as β and σ^2. Given the highly specific form of the covariance matrix, there is no simple conjugate update, so τ will be sampled with Metropolis step. The proposal will be $\tau^{(n)} \sim \mathsf{Un}(\tau^{(n-1)} - \delta, \tau^{(n-1)} + \delta)$.

2.5 Results

In this section we have results from the two models discussed above. We begin with our OLS comparison.

2.5.1 Comparison of OLS and Bayesian Model

Suppose we run the regression

$$\log V_i = \beta_0 + \beta_1 \texttt{Population} + \beta_2 \texttt{Population}^2 + \beta_4 \texttt{Bandwidth} + \epsilon_i. \tag{2.1}$$

There are many other potential covariates that are used in the literature, including the median income of the license area, the population density, and variables related to the bidder. Some of these are certainly important covariates and their omission

could create omitted variable bias. These are not problems we are concerned with here because our interest is in comparing two methods and these problems should affect both similarly. We have applied a log transformation to the response to reduce the skewness of the data. We have also restricted our estimates to *winning* bids, as is commonly done in the literature.

The OLS estimates for this model are given in Table 2.3. All coefficients except β_4 are statistically significant. Population has been scaled by 1000, so the estimate β_2 represents the expected increase in $\log V$ for an increase of 1000 people in the license area. There are $n = 1090$ observations and the adjusted R^2 is 0.24.

Table 2.3: OLS estimates for Equation 2.1 for winning bids.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	13.5981	0.2439	55.76	0.0000
Population	0.0003	0.0000	17.53	0.0000
Pop2	-0.0009	0.0001	-15.01	0.0000
Bandwidth	0.0292	0.0216	1.35	0.1776

The results of the Bayesian model are summarized in Figure 2.7. The histogram represents 1000 posterior samples for the regression coefficients. The black horizontal line is 95% confidence interval for the estimates in Table 2.3. One interesting feature is that for β_1 and β_4 the posterior distributions coincide very closely with the OLS estimates. Note that these are the covariates with the least variance (see Table 2.2 for the distribution of the bandwidth variable). For population and population squared, we clearly see that the OLS estimates are less than the Bayesian estimates. For β_3 in particular there is no overlap between the 95% confidence interval and the Bayesian posterior. The posterior variance is so small compared to the OLS variance that the posterior appears concentrated on a point. On the other hand, it is clear that β_3 is very small, so the estimates may be practically equivalent.

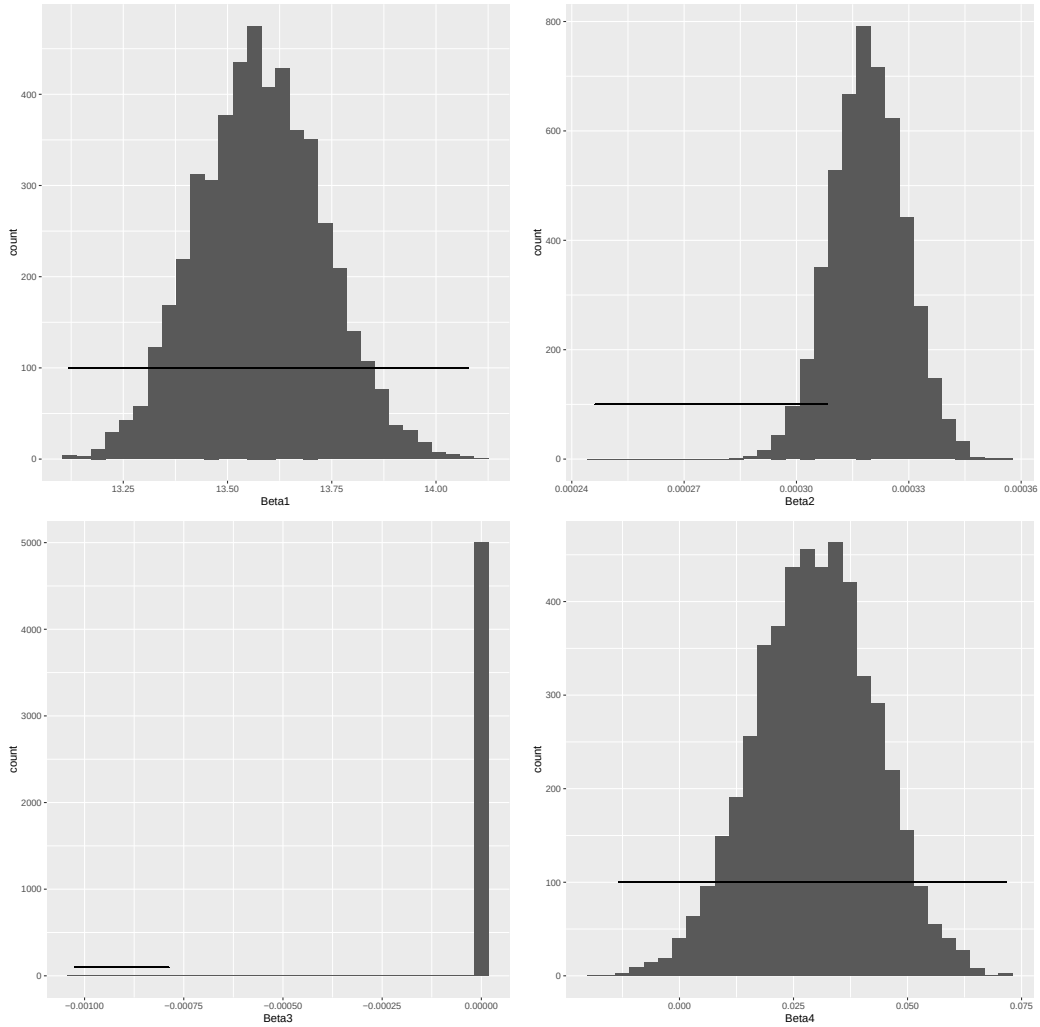

FIGURE 2.7: Posteriors for (clockwise from top left) $\beta_1, \beta_2, \beta_4$, and β_3.

2.5.2 Spatial Model

In the spatial model, we consider the following regression model:

$$\log V_i = \beta_0 + \beta_1 \texttt{Population} + \beta_2 \texttt{Population}^2 + \beta_3 \texttt{Bandwidth} + \tag{2.2}$$
$$\beta_4 \log \texttt{Eligibility} + \epsilon_i$$

with a correlation structure for each firm given by $(\mathbf{I}_{n_i} + \tau T_i)$, where T_i the matrix indicating which of the licenses that firm i bid on share a boundary.

Switching from a regression model with a diagonal covariance matrix to one that allows dependence represents a tradeoff. The benefit is that we can create a

more realistic model that better reflects how we believe bidders make choices. The problem is that the computational complexity increases drastically. As an example, Verizon Wireless bid on 615 licenses and AT&T bid on 549 licenses. Therefore we must sample from a 500-dimensional *truncated* multivariate normal. The truncation creates problems, especially in high dimensions. In low dimensions, samples from a truncated normal can be drawn using rejection sampling. When the dimension exceeds two or three, this becomes very inefficient and a Gibbs' sampler is used. If 100 iterations are used as a burn-in period, then drawing valuations for Verizon alone requires 62,115 samples to be drawn. Imputing valuations for all firms once takes several minutes, meaning that a Gibbs sampler takes several days to complete even a modest number of iterations. Our solution is to restrict our attention to bidders who placed bids on no more than ten licenses. From an economic perspective, this is actually somewhat appealing, as many of the firms that bid on a small number of licenses are local communications companies who operate in a few cities or counties. This might cause them to have more highly correlated valuations for adjacent regions.

When restricted to companies that placed bids on no more than ten licenses, the OLS estimates for Model 2.2 are given in Table 2.4. Note that the spatial structure is *not* imposed for the OLS estimates. There are 453 observations and the adjusted R^2 is 0.354.

Table 2.4: OLS estimates for Model 2.2.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.8885	0.5522	14.29	0.0000
Population	2.0564	0.2160	9.52	0.0000
Bandwidth	0.0569	0.0336	1.69	0.0912
Pop2	-74503.9759	8471.8853	-8.79	0.0000
Log_Elig	0.2502	0.0234	10.71	0.0000

The posterior estimates are given in Figure 2.8. The relationship between the posterior and the OLS estimates is similar to that shown in Figure 2.7. The posterior

distribution for the intercept β_1 is similar to that given in Figure 2.7, but the OLS estimate is noticeably smaller. For the other parameters, the posterior is much more concentrated than the OLS confidence interval, which is interesting because we are adding uncertainty by imputing new valuations at every iteration. The most interesting plot is the posterior for τ in the bottom right. From a uniform prior, we get a highly concentrated posterior, centered around $\tau = 0.35$. This means that there is strong evidence of a positive correlation in valuations between adjacent license regions.

2.6 Analysis of Number of Bidders

William Vickery essentially founded the field of auction theory with his seminal 1961 paper *Counterspeculation, Auctions, and Competitive Sealed Tenders* (Vickery (1961)). In this paper, he showed that for a single round, sealed bid first-price auction, there was a symmetric Nash equilibrium where each bidder with valuation independent valuations $V_i \sim \mathsf{Un}(0, 1)$ placed bid B_i given by

$$B_i = \frac{n-1}{n} V_i,$$

where n is the number of bidders. This means that as the number of bidders increases, the winning bidder's surplus, $V_i - B_i$, decreases to zero as $n \to \infty$. We will consider the ratio B_i/V_i, which should approach one as $n \to \infty$.

The specifics of this result are not relevant for our context: the auction is not a single round, the valuations are not uniformly distributed, and the valuations are not independent. The general idea, however, that a the number of bidders increase, the winning bid will be closer to the valuation, as indicated in Figure 2.9.

The goal of this section is to attempt to verify this result empirically. Specifically, we wish to show that an increase in the number of bidders is associated with an increase in the expected value of the ratio B_i/V_i for the winning bidders. Specifically,

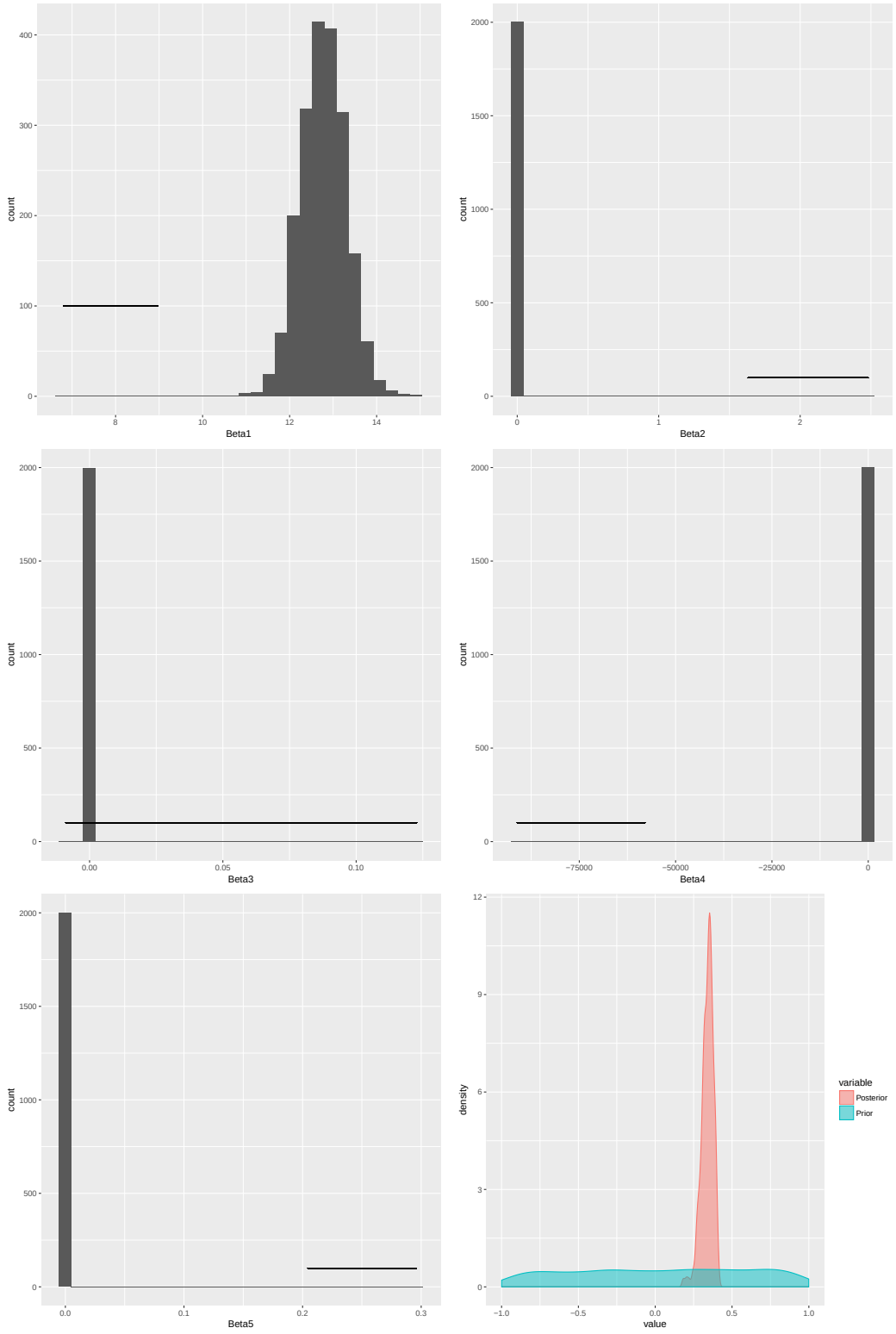

FIGURE 2.8: Posterior Estimates for Mode 2.2 with OLS 95% Confidence Intervals

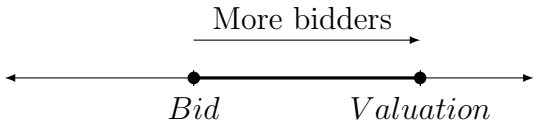

FIGURE 2.9: An increase in the number of bidders forces each bidder to bid closer to their valuation, reducing surplus.

we restrict our data to the winning bids and consider the posteriors generated for these winning bids. For a large number of samples from the posterior, we compute the ratio B_i/V_{ij} for $j = 1, \ldots, S$. This creates a distribution for the ratio. We sample from this distribution for each winner and run a simple OLS regression of the winner's ratio on the number of bidders for the license, which we observe in the data. By repeatedly sampling, we can get a distribution for the regression coefficients and attempt to draw conclusions.

2.6.1 Examples of Valuation Posteriors

The posterior for the valuation is a truncated normal. For winning bids, the posterior is a truncated normal that is unbounded above. The truncation point is the value of the winning bid. If there is significant density above the truncation point, then the model is concluding that the valuation could be much larger than the winning bid. When there is little density above the bid, then the model believes that the bidder has been forced to bid close to their valuation. These circumstances are illustrated in Figure 2.10. The situations shown in Figure 2.10 are actually observed in the data. Actual posteriors for winning bids in AT&T are shown in 2.11. The CMA56 license covers Northeast Pennsylvania and CMA33 covers San Antonio, TX. The bidding for CMA33 involved seventeen bidders (the largest number in Auction 73), while the bidding for CMA33 included only three bidders. Note how for CMA56 we see the mode of the truncated normal, indicating that the model predicts that AT&T's valuation could be significantly higher than the winning bid. For CMA33,

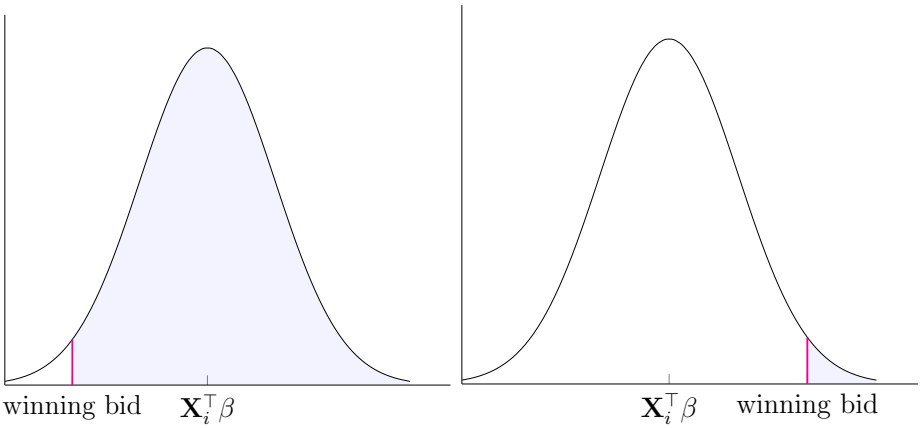

FIGURE 2.10: Sample posteriors for an "easy" win (left) and a "hard" win (right).

on the other hand, we clearly see that the truncation point is in the upper tail of the normal, indicating that AT&T won the license with a bid closer to the valuation.

The distribution of the ratio B_i/V_i for the licenses shown in Figure 2.11 is shown in Figure 2.12. As expected, the mode for CMA56 is smaller than the mode for CMA33, indicating that AT&T would have been willing to bid significantly higher for CMA56. One feature of note is the support of the distributions. Both distributions have support from zero to one. This reflects the fact that there is significant uncertainty about the valuation. The uncertainty is undesirable, but also shows that our model does not impose strict assumptions on the valuations.

2.6.2 Number of Bidders and Surplus

Once we construct the distributions of B_i/V_i, we can sample a value from each winning bidder's distribution for each license and compute the OLS estimates of

$$\frac{B_i}{V_i} = \beta_0 + \beta_1 n_i + \epsilon_i, \tag{2.3}$$

where n_i is the number of bidders who placed bids on the license. By repeated sampling and estimation, we can get a distribution of estimates for (β_0, β_1). This distribution is presented as regression lines in Figure 2.13. We are interested in determining whether the slope is positive. Examining Figure 2.13, we see that most lines

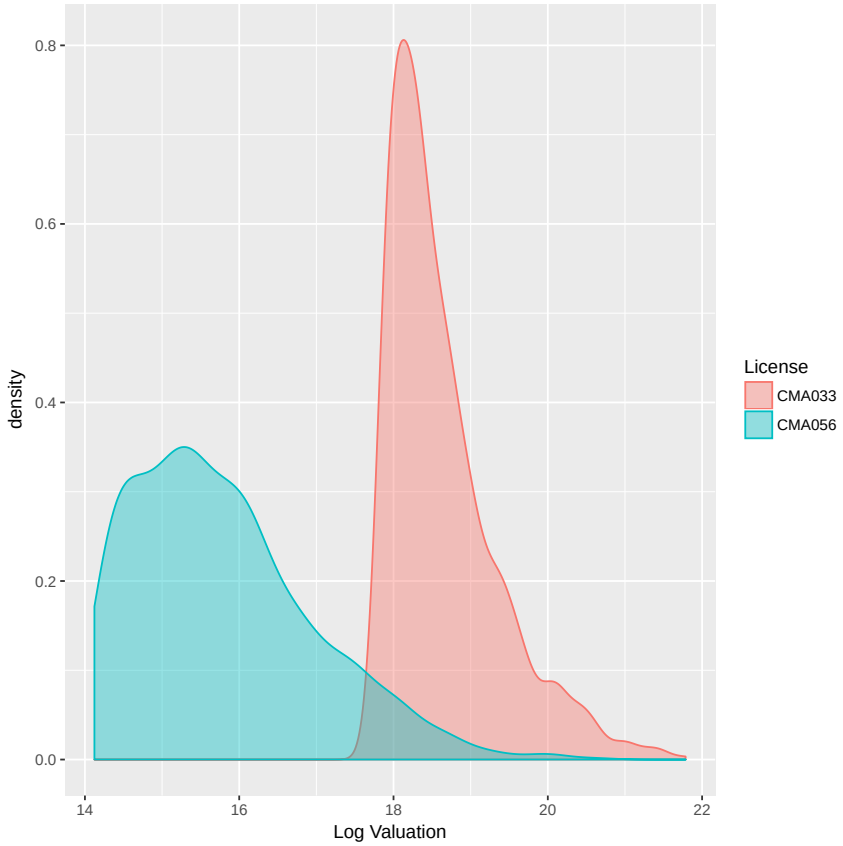

FIGURE 2.11: Posteriors for AT&T's valuations of licenses CMA56 and CMA33.

have a positive slope, but not all. Furthermore, the regression lines are concentrated for certain values of n (around $n = 5$), but very dispersed for large values of n. The variance in the value of the regression line is explained by Figure 2.14. The high large number of observations with n around five means that $\mathsf{E}\left[B_i/V_i \mid n_i\right]$ is precisely estimated for small n but highly variable for large n.

To test whether we expect more bidders to increase competition and decrease the winner's surplus, we must look at the distribution of estimates for β_1, which is shown in Figure 2.15. The empirical probability that the estimate is positive is $\mathsf{P}\left[\beta_1 > 0\right] = 0.994$, which is very strong evidence. In other words, using our model we can empirically verify a well established and intuitive result from auction theory.

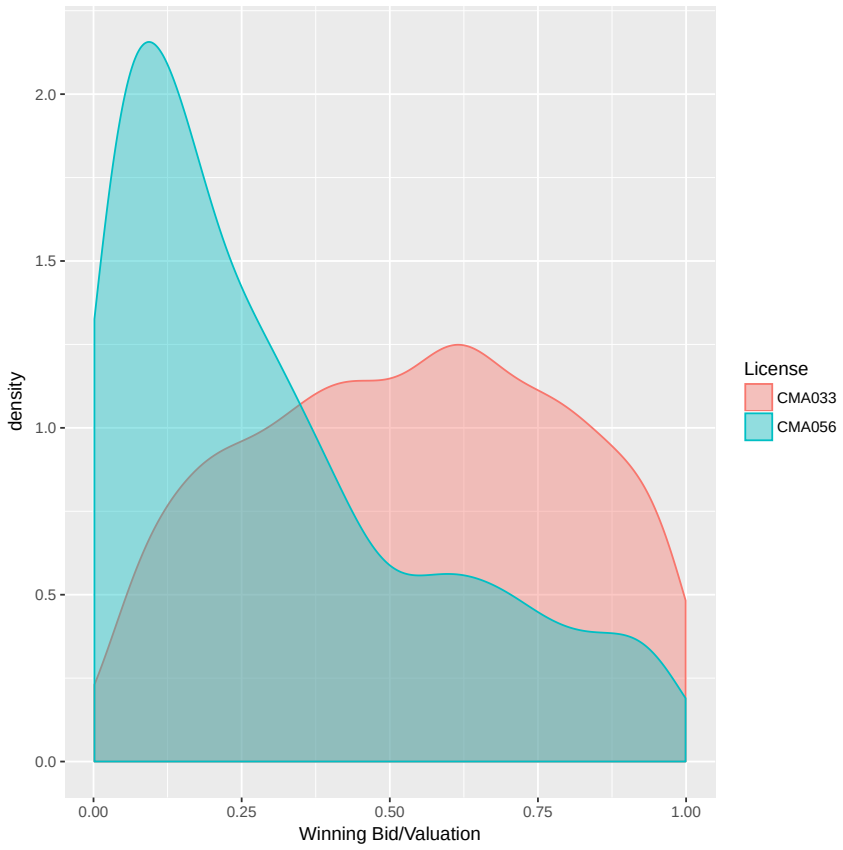

FIGURE 2.12: Distribution of ratio B_i/V_i with V_i sampled from posterior for licenses CMA56 and CMA33.

2.7 Conclusion

In this chapter we analyzed auctions for wireless spectrum held by the FCC. A simple analysis of the structure of the auctions gives us bounds on missing values that we do not observe. Examining the data, we have reason to believe that our theoretical analysis is appropriate. When we compare the results of OLS estimation with a new Bayesian model, we find that some parameters are estimates are similar across models, while others are quite different. We also propose a model that considers whether two licenses are adjacent and estimates the correlation between adjacent license regions. We find evidence of a moderate correlation between the bidders' valuations of adjacent license regions. Finally, we empirically verify the result from

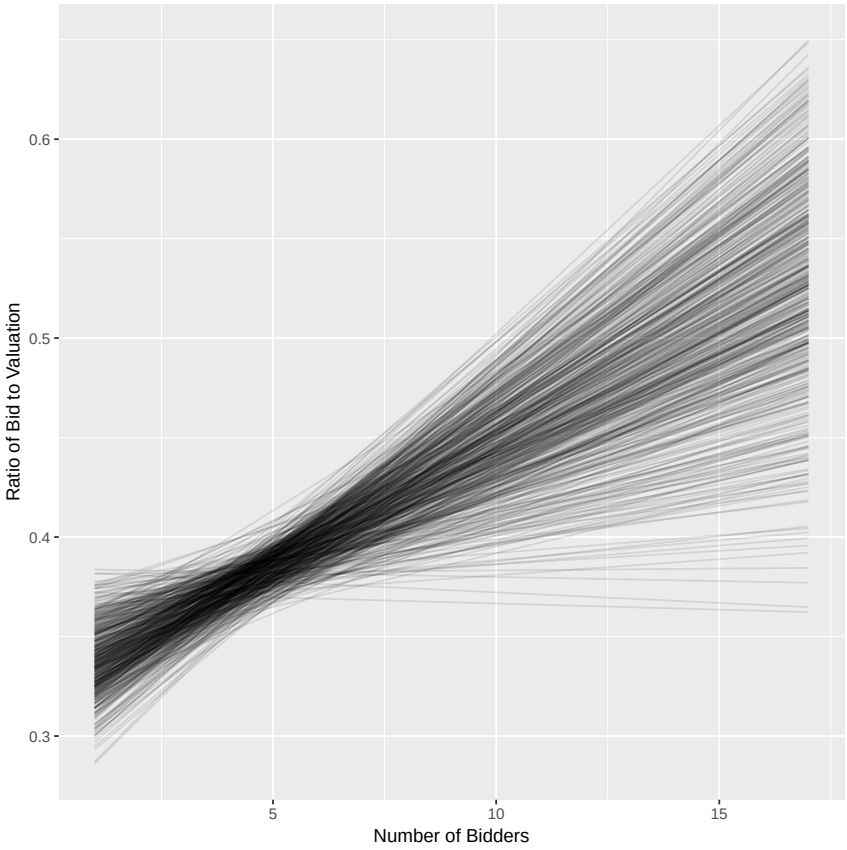

FIGURE 2.13: Distribution of OLS regression lines.

auction theory that more bidders increase competition and force bidders to bid closer to their valuations.

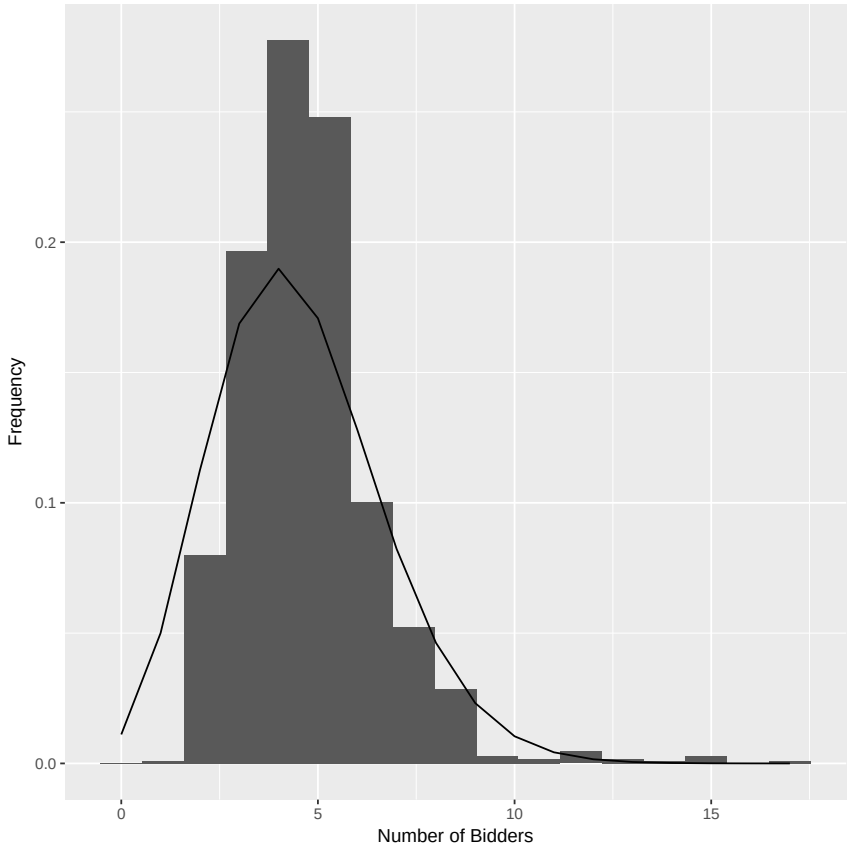

FIGURE 2.14: Histogram of number of bidders per license with a Poisson density (with mean at MLE) imposed on top.

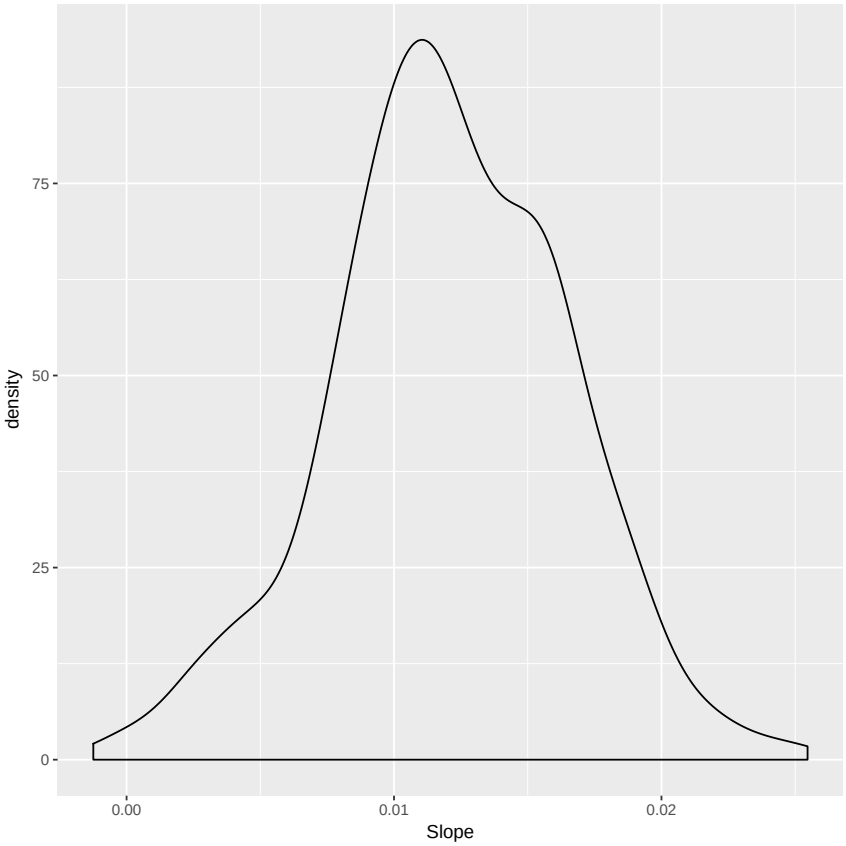

FIGURE 2.15: Distribution of estimates for β_1.

Appendix A

Full Conditionals

First we augment with a categorical variable Z for each sample.

$$Z_{ij} \mid \theta_i \sim \mathsf{Mult}(1, \theta_i)$$

$$X_{ij} \mid Z_{ij} \sim \mathsf{No}(\mu_{z_{ij}}, 1).$$

Now we model θ_i as arising from stick breaking a vector ψ_i. Let π_{SB} denote the stick breaking function, so $\theta_i = \pi_{SB}(\psi_i)$. Then the model is

$$\psi_i \sim \mathsf{No}(0, \sigma^2 I_n)$$

$$\theta_i = \pi_S B(\psi_i)$$

$$Z_{ij} \mid \theta_i \sim \mathsf{Cat}(1, \theta_i)$$

$$X_{ij} \mid Z_{ij} \sim \mathsf{No}(\mu_{z_{ij}}, 1).$$

Full conditionals:

$$\begin{aligned} p(\psi \mid \Sigma, \theta, \omega, Z, X) &\propto p(\psi \mid \Sigma, \omega, Z) \\ &\propto \mathsf{No}(\kappa(c), \Omega^{-1})\mathsf{No}(\psi \mid 0, \Sigma) \\ &\propto \mathsf{No}(\psi \mid \tilde{\mu}, \tilde{\Sigma}), \end{aligned}$$

where

$$\begin{aligned}
\Omega &= \operatorname{diag}(\omega) \\
\tilde{\Sigma} &= \left[\Omega + \Sigma^{-1}\right]^{-1} \\
\tilde{\mu} &= \tilde{\Sigma}\left[\kappa(c) + \Sigma^{-1}\mu\right] \\
c_t &= \sum_{i=1}^{n} \mathbf{1}[z_i = t] \\
\kappa(c) &= c - N(c)/2 \\
N(c) &= (N_1, \ldots, N_k) \\
&= \begin{cases} N_1 = N = \sum_{i=1}^{k} x_i \\ N_i = N - \sum_{j<i} x_j. \end{cases}
\end{aligned}$$

For ω, the Polya-Gamma variable, we have

$$p(\omega \mid Z, \psi) = \mathsf{PG}(N(Z), \psi),$$

which is the standard Polya-Gamma update. For the categorical variable Z_i the update is

$$\mathsf{P}\left[z_{ij} = \ell \mid \theta, x_{ij}\right] \propto \theta_{i\ell} f(x_{ij} \mid \mu_\ell),$$

where $f(x_{ij} \mid \mu_{z_{ij}})$ is a normal density with mean μ_ℓ evaluated at the observed value x_{ij}. In this case, it is easy to find the normalizing constant:

$$\mathsf{P}\left[z_{ij} = \ell \mid \theta, x_{ij}\right] = \frac{\theta_{i\ell} f(x_{ij} \mid \mu_\ell)}{\sum_{n=1}^{k} \theta_{in} f(x_{ij} \mid \mu_n}$$

Bibliography

Bohlin, E., Madden, G., and Morey, A. (2010), "An Econometric Analysis of 3G Auction Spectrum Valuations," *EUI RSCAS, Florence School of Regulation*.

Caruso, V. C., Mohl, J. T., Glynn, C., Lee, J., Willett, S. M., Zaman, A., Estrada, R., Tokdar, S., and Groh, J. M. (2017), "Evidence for time division multiplexing of multiple simultaneous items in a sensory coding bottleneck," *bioRxiv*.

Connolly, M., Sa, N., Roark, C., Zaman, A., and Trivedi, A. (2017), "The Evolution of U.S. Spectrum Values Over Time," .

Glynn, C. (2016), "Advances in Dynamic Modeling and Computation for Count Data," Ph.D. thesis, Duke University.

Liu, X. and Daniels, M. J. (2006), "A New Algorithm for Simulating a Correlation Matrix Based on Parameter Expansion and Reparameterization," *Journal of Computational and Graphical Statistics*, 15, 897–914.

Polson, N. G., Scott, J. G., and Windle, J. (2013), "Bayesian Inference for Logistic Models Using Pólya–Gamma Latent Variables," *Journal of the American Statistical Association*, 108, 1339–1349.

Vickery, W. (1961), "Counterspeculation, Auctions, and Competitive Sealed Tenders," *Journal of Finance*, 16, 8–37.